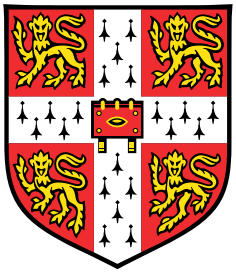
Tissue-specific adaptations of cell types



Tomás Pires de Carvalho Gomes

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

King's College

January 2020

To the bigger picture!

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Contributions. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations.

Tomás Pires de Carvalho Gomes
January 2020

Acknowledgements

I would need an additional very long chapter to fully and fairly acknowledge each and every person that contributed to this endeavour.

I would like to acknowledge my supervisor, Sarah Teichmann, for the opportunity to pursue this PhD. It has been a long and winding journey, and I could not have done it without her guidance. Sarah was always ready to provide valuable, original insights on the problems and questions I had, teaching me so much more than I have ever expected. I fell truly fortunate to have her as a mentor that I can rely on.

Besides being a brilliant scientist, one of Sarah's greatest achievements has been to put together and maintain a highly collaborative lab, full of outstanding people always ready to inspire and help. It is hard to find someone who has not had an impact on my progress and research, but I would like to name a few. To Ricardo Miragaia, the other half of the Portuguese dynamic duo, for the relentless understanding and support provided. Through good and bad times, we had a lot of fun together, and I would not have it any other way. I also want to acknowledge my "Jedi master" Roser Vento, for the interesting, deep discussions and crucial encouragement throughout my PhD. I am lucky to have such a brilliant and dedicated friend. To Tzachi Hagai, who is a constant source of inspiration, enthusiasm and skepticism. He has taught me to question everything, and through better knowing my limitations make the most of them. Valentine Svensson was one of the people that I learned the most from. Likely the most talented young scientist I have ever met, he was crucial in my understanding of single-cell computational methods, and always a fun company to be around. Raghd Rostom, my "PhD sister", whose funny antics and incisive comments have helped keep my head grounded. To Kylie James, for the uplifting conversations about life, they have always left me with a feeling of hope for the future. And to Rasa Elmentaite, for her inspiring dedication and contagious desire to learn and ask questions.

I am greatful to the whole ENLIGH-TEN consortium, which funded my PhD and provided me with the opportunity to meet several other early stage researchers and

their supervisors. The support from this network has been crucial for my research, and gave me a true sense of what it is like to do science across borders.

I could not have started this PhD without the assitance of my former advisors Maria do Carmo Fonseca and Ana Rita Grosso, as well as collaborators Nick Proudfoot and Taka Nojima. They were crucial in starting my career in bioinformatics and genomics, and are overall great scientists and supportive mentors.

I would also like to thank all my friends, in particular to Tiago Pires, who despite the distance always gave me an escape from the hectic PhD life. I would also like to thank Mariana, Elsa, Gonçalo, and Joana for the fun times we had in Cambridge and with the Cambridge University Portuguese Speakers' Society. And to Gianmarco Raddi, who I could always count on to keep my spirits up.

I am deeply grateful to all my family, my father José Carlos, my mother Ana Luísa, my brother Miguel, my aunt Fernanda, and my grandmother Maria do Carmo.They have been relentless in their assistance and understanding, and it is thanks to them that I got the opportunity to work on what I love. I also want to posthumously thank my grandfather Fernando for all the inspiration and all he taught me.

Finally, and most importantly, I want to thank my partner Hajrabibi Ali. I am deeply indebted to her for her encouragement, patience and self-sacrifice, that ultimately help me pursue this degree. I am hoping that I can ever repay her for her time and help in keeping me focused and reminding me of the greater picture.

Contributions

The multidisciplinary nature of the studies here presented required the valuable contributions of my collaborators. This will be further detailed at the start of each chapter, but will also be here summarised.

- In Chapter 2, the original experiments were designed by Ricardo J Miragaia, who also assisted in data interpretation.
- In Chapter 3, the original concept was conceived together with Valentine Svensson.

Abstract

Cells are the building blocks of life, forming the vast diversity of tissues and organisms in Nature. Across these, common cellular morphologies and functions have been identified. High-throughput, multifactorial profiling of cells has grown exponentially in recent years with the advent of single-cell RNA-sequencing (scRNA-seq), increasingly unravelling cell diversity. Nonetheless, it is not yet known how different environments affect cellular phenotypes.

The work presented on this Thesis reports on the transcriptional variation of cell types across tissues, by use of single-cell RNA-sequencing. This technology, developed in the last 10 years, has greatly impacted our ability to distinguish cellular heterogeneity by their gene expression in various tissues or conditions.

Chapter 1 outlines the impact of single-cell RNA-sequencing in cell biology, presenting the technology as the natural progression of lower throughput or low-resolution methods. The chapter then shows how cellular heterogeneity can be deconstructed by analysing this type of genomics data. It then expands on how individual datasets can be used to build models of cell type identity for automatic annotation, ultimately outlining the need to create a global cell type census of a whole organism. A cell compendium like this should be useful for automatic annotation, as well as to obtain a cross-tissue integrative overview of cell identity.

The same chapter also delves into the topic of heterogeneity in immune cells. Due to the evolutionary pressure they are subject to and ubiquitous nature across the organism, these are some of the most diverse cell types in multicellular organisms. Chapter 2 presents a deconstruction of T-regulatory cells' phenotypes in different mouse and human tissues using single-cell RNA-sequencing. The analysis in this chapter will show how these cells are structured in subpopulations, and how they adapt when migrating between lymphoid and non-lymphoid tissues. It will also assess the conservation of gene expression programmes for the same populations between mouse and human.

The creation of a global cell type reference is an endeavour that can facilitate analysis of new data, and reveal novel insights about cell and tissue biology. Several

datasets have now been produced, and a method that can efficiently integrate them and prepare them for use as a reference is necessary. Chapter 3 details the development of such method, exploring its strengths and how it can be improved, in a mouse dataset. Chapter 4 then applies this pipeline to a collection of human data, and shows how cell types relate across tissues, as well as how the human reference can be used in a practical case.

Lastly, Chapter 5 summarises all chapters, providing an overview on how single-cell sequencing has changed what we know about tissue biology, and how listing cell types and compiling them as a functional reference can help future developments in life sciences.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

APC Antigen-Presenting Cell

ARD Automatic Relevance Determination

BGPLVM Bayesian Gaussian Process Latent Variable Modelling

bLN brachial Lymph Nodes

CITE-seq Cellular Indexing of Transcriptomes and Epitopes by sequencing

cTreg central Treg (cells)

DE Differentially Expressed (genes)

EGFP Enhanced Green Fluorescent Protein

ERCC External RNA Controls Consortium

eTreg effector Treg (cells)

FACS Fluorescence-Activated Cell Sorting

GRCh Genome Reference Consortium human

GRCm Genome Reference Consortium mouse

HCA Human Cell Atlas

iNKT invariant Natural Killer T (cells)

LN Lymph Nodes

LT Lymphoid Tissues

LV Latent Variable

MHC Major Histocompatibility Complex

mLN mesenteric Lymph Nodes

MRD-BGPLVM Manifold Relevnce Determination-BGPLVM

NLT Non Lymphoid Tissue

NMF Non-negative Matrix Factorization

oNMF orthogonal Non-negative Matrix Factorization

PBS Phosphate Buffer Saline

PCA Principal Component Analysis

QC Quality Control

RNA Ribonucleic acid

scATAC-seq single-cell Assay for Transposase-Accessible Chromatin sequencing

scRNA-seq Single-cell RNA sequencing

SGD Stochastic Gradient Descent

SJ split-join (distance)

SS2 Smart-seq2

SVM Support Vector Machine

TCR T Cell Receptor

Tfh T follicular helper (cells)

Th T-helper (cells)

Tmem T-memory (cells)

Treg T-regulatory (cells)

tSNE t-Distributed Stochastic Neighbor Embedding

VAE Variational Autoencoder

VAT Visceral Adipose Tissue

Chapter 1

Cellular identity in the genomics era

Cell biologists have attempted, from the inception of the discipline, to categorize the extensive variability of cells that are found in Nature. This endeavour is hampered by the intrinsic complexity of cells, which associated to their small size and sensitivity to the surrounding environment, makes cellular phenotypes hard to probe in an integrated and comprehensive way. The last decade however has seen extraordinary improvements in the detail to which molecules can be assayed in individual cells. Single-cell RNA-sequencing (scRNA-seq) has for the first time provided an unbiased, transcriptome-wide census of RNA molecules for one cell at a time. By acquiring the transcriptome of large numbers of cells, we can group them by their gene expression programmes - a proxy for their function - and thus define their cell identity. The definition of this cell type identity from the massive amounts of transcriptome data produced in recent years has required the continuous adoption of new computational and analytical methodologies.

This chapter provides an introduction to the definition of cell types. It will show how more recently developed experimental and computational approaches are shaping our understanding of how cells are categorized.

1.1 Cell type discovery and definition

The term "cell" was coined by Robert Hooke in the 17th century to describe the empty cell walls he observed in cork samples through his microscope (Hooke, 1667). This observation was complemented some years later, when Antonie van Leeuwenhoek first observed live unicellular organisms and other cells with a microscope composed of more powerful lenses (Mazzarello, 1999). Research and observations in the

following 200 years led to the formulation of cell theory. Its first tenet was introduced by Schleiden and Schwann, and states that all living structures are composed of cells or their byproducts (Schwann, 1847). The theory was later complemented by Robert Remak, Rudolf Virchow, and Albert Kölliker to include the postulate that all cells are derived from other cells (in the latin formulation popularized by Virchow, *omnis cellula e cellula*).

These early studies looked at a variety of sources to unveil different types of cells. Leeuwenhoek reported observations from blood, brain, muscle and semen (Leeuwenhoeck M, 1674; Leeuwenhoek Antoni Van, 1677). Subsequent developments of microscopy techniques led to improved imaging of a variety of tissues and the cells that compose them. For the first centuries of cell biology, microscopy was the method of choice to identify cell types. While this was mostly due to the relatively reduced knowledge of cellular biochemistry, it was immediately apparent that morphology was intrinsically tied to cellular function. The most illustrative example of this is the neuron, whose unique structure was only unravelled after subsequent improvements in tissue preparation and staining, as well as increases in resolution and development of electron microscopy (Mazzarello, 1999). Microscopy was also important in understanding where cell types come from by mapping their developmental origin. The three germ layers - endoderm, mesoderm, ectoderm - were identified in the 19th century, and was postulated that each of them would give rise to different sets of tissues (Collins and Billett, 1995). Developmental studies have since had a central role in defining cell lineages, and thus how cell types are related. Advances in microscopy were also crucial to the identification of organelles. While larger structures, like nuclei, are still identifiable with simpler microscopes (Brown, 1866), others required improved resolution and staining or preparation to be identified (Golgi and Lipsky, 1989). Other advancements in microscopy like live-cell imaging or super resolution microscopy are constantly perfected to expand the boundaries of cellular functional characterization.

Advances in biochemistry and molecular biology revealed that most organic molecules that compose cells are directly responsible for their function. Proteins are responsible for most cellular functions, being involved in enzymatic reactions, signalling and regulatory pathways or structural components. They became a prime target for cellular phenotyping with the development of immunostaining (Coons et al., 1941), whereby an antibody that specifically targets a certain protein is usually tagged with a fluorophore. Immunostaining can identify protein expression in tissue slices, and the use of different fluorophores allows for the imaging of cells expressing

multiple proteins. The usefulness of immunostaining became especially apparent when it was combined with high-throughput microfluidics methods and used for fluorescence-activated cell sorting (FACS) (Bonner et al., 1972). This introduced the first high-throughput studies on molecular phenotyping of cells, and sorting allowed cell function to be probed in parallel (Julius et al., 1972). More recently, mass cytometry has allowed for a further expansion of the repertoire of proteins assayed (Bandura et al., 2009; Di Palma and Bodenmiller, 2015). This technique, while destructive, has also been combined with tissue imaging, adding a spatial component to the cell populations examined (Chang et al., 2017).

The identification and classification of cell types is dependent on their function. Function is deeply related to cellular morphology (Prasad and Alizadeh, 2019), and both are ultimately a consequence of the molecular pathways shaping them. Additionally, even though recent advances permit high throughput cell sorting through imaging (Nitta et al., 2018), the limited resolution hinders the identification of finer details of cell and organelle shape, which are frequently more informative of cellular activity. Cell sorting with fluorescent antibodies and mass cytometry can reveal more details on the molecules underlying cellular behaviour, but they are targeted approaches that depend on prior knowledge of the effector molecules. The more recent attempts at defining cell identity have therefore relied on the unbiased, high-throughput character of single-cell RNA-sequencing methods.

1.2 Defining cell types using scRNA-seq

Methods to sequence the transcriptome of individual cells started to be developed shortly after the advent of RNA-seq (Mortazavi et al., 2008; Tang et al., 2009). This early development was pushed not by a need to define the molecular makeup of the unit of life, but rather to allow transcriptomic studies to be performed in low-input samples. Nonetheless, this seminal work still sparked the improvements that occurred in the decade that followed (Svensson et al., 2018) (Figure 1.1).

Initial developments focused on increasing sensitivity, since the original scRNA-seq protocol was performed on cells from very early developmental stages, which are larger and contain more RNA than most differentiated cell types. Different methodologies quantified gene expression by sequencing distinct transcript segments (either the 5' or the 3' end, or the full transctipt) (Hashimshony et al., 2012; Islam et al., 2011; Picelli et al., 2014; Ramsköld et al., 2012). The idea of multiplexed scRNA-seq also started gaining traction with the use of multi-well plates or molecular

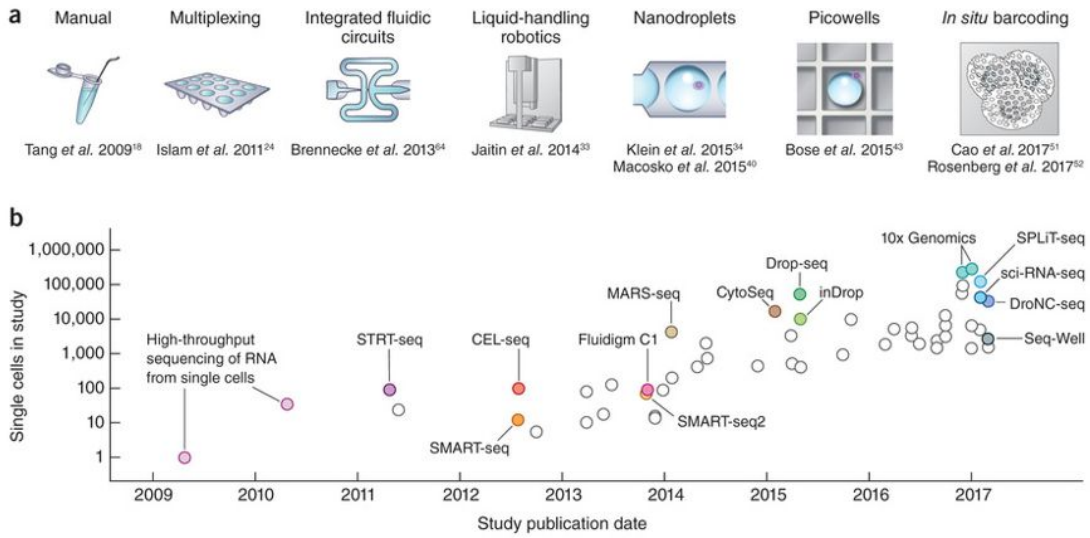


Fig. 1.1: **Timeline of scRNA-seq technology development**
(A) Key technologies that have allowed jumps in experimental scale. A jump to ~100 cells was enabled by sample multiplexing, and then a jump to ~1,000 cells was achieved by large-scale studies using integrated fluidic circuits, followed by a jump to several thousands of cells with liquid-handling robotics. Further orders-of-magnitude increases bringing the number of cells assayed into the tens of thousands were enabled by random capture technologies using nanodroplets and picowell technologies. Recent studies have used in situ barcoding to inexpensively reach the next order of magnitude of hundreds of thousands of cells. **(B)** Cell numbers reported in representative publications by publication date. Key technologies are indicated. *Original figure published in (Svensson et al., 2018).*

barcodes for cells. The company Fluidigm eventually introduced the first commercially available microfluidics chips (called the "Fluidigm C1 system") for miniaturized cell isolation, RNA extraction and reverse transcription (Brennecke et al., 2013). It is from this point that increased cell capture becomes the major technological driver (and has gained even great importance as discussed in Section 1.3). The major contributors to this have been nanodroplet-based technologies, that have put the number of profiled cells per dataset in the range of 10.000 to 100.000 (Klein et al., 2015; Macosko et al., 2015). The importance of this increase in throughput has been demonstrated by Shekar and colleagues (Shekhar et al., 2016), where they demonstrate that a Drop-seq dataset of approximately 25.000 cells sequenced at low depth could identify more *bona fide* cell types and subtypes than a smaller, more deeply sequenced Smart-seq2 dataset. Currently, most single-cell RNA-seq datasets use droplet-based technologies, chiefly the protocols designed for the Chromium

instrument by 10x Genomics (Zheng et al., 2017), which have a higher sensitivity to detect different transcripts. Other more recent methods have followed the trend of increase in cell throughput by using multiplexed barcoding, which allows for different samples to be combined and reducing sample processing costs, reaching 10^5-10^6 cells for less than $0.01 per cell (Cao et al., 2019a; Rosenberg et al., 2018). A list of the most up-to-date scRNA-seq methods can be found in Table 1.1.

Table 1.1: Current methods for single-cell RNA-sequencing

Method Name	**Reference**
Fluidigm C1	(Brennecke et al., 2013)
Smart-seq2	(Picelli et al., 2014)
Drop-seq	(Macosko et al., 2015)
inDrop	(Klein et al., 2015)
CEL-seq2	(Hashimshony et al., 2016)
Chromium	(Zheng et al., 2017)
ICELL8	(Goldstein et al., 2017)
Quartz-seq2	(Sasagawa et al., 2018)
mcSCRB-seq	(Bagnoli et al., 2018)
SPLiT-seq	(Rosenberg et al., 2018)
MARS-seq2	(Keren-Shaul et al., 2019)
sciRNA-seq3	(Cao et al., 2019a)
Seq-Well S^3	(Hughes et al., 2019)

The exponential developments in single-cell sequencing technologies were accompanied by essential computational developments to analyse the resulting data. From a cell type discovery perspective, the key methods are clustering and pseudotime analysis (Rostom et al., 2017), which assign to cells a discrete or a continuous label, respectively. These are of course dependent of the upstream processing steps of normalisation, feature selection and dimensionality reduction, as well as often used batch correction methods (Luecken and Theis, 2019). Most of these analysis steps are available in accessible software toolkits (Butler et al., 2018; McCarthy et al., 2017; Wolf et al., 2018).

With clustering, the goal is to identify discrete cell populations. The most widely used methods for clustering are the louvain and leiden community detection algorithms (Blondel et al., 2008; Traag et al., 2019). These populations are commonly considered an approximation of the cell types present in a sample of dataset, often justified by examining the presence of known markers for known cell types across clusters. Further application of differential expression methods (extensively benchmarked in (Soneson and Robinson, 2018)) between clusters can identify other potentially novel genes that are, within that context, unique to that population. This can be used to characterise newly discovered populations (Montoro et al., 2018; Shekhar

et al., 2016; Villani et al., 2017) and to identify new markers that can be used to isolate or understand known cell types (Bjorklund et al., 2016; Shulse et al., 2019; Vento-Tormo et al., 2018).

Pseudotime analysis consists on describing a set of cells from a continuous perspective. The name derives from the original application to obtain a dimensionless temporal trajectory from time course scRNA-seq data (Trapnell et al., 2014). There are several methods to perform this analysis (exhaustively reviewed in (Saelens et al., 2019)), all with the goal of defining a latent variable from the data along which a biological process, reflected in gene expression, is changing. Pseudotime is especially useful to study response to stimuli (Lönnberg et al., 2017; Trapnell et al., 2014) and developmental trajectories (Cao et al., 2019a; Watcham et al., 2019), but has also been used to model changes to cellular spatial distribution (Scialdone et al., 2016). These methods can differ in the way they model biological trajectories, with some explicitly allowing for branched trajectories. This is of special importance in development, where the goal is usually understanding which daughter cell types share progenitors. The direction of differentiation is usually just assumed according to previous knowledge and of the experimental conditions. This is not completely possible in all situations, yet can be inferred from expression data. By considering RNA kinetics, and using the quantification of spliced/unspliced reads, the current and future (i.e. still circumscribed to the nucleus) transcriptomic states can be untangled as a "velocity" vector (Manno et al., 2018). In differentiation trajectories, cell types are therefore usually defined as the endpoints, with the cells in between forming more transient cell states, along which gene expression is dynamically adjusting to the final cellular identity. It should be noted that this "cell type vs cell state" nomenclature is context-dependent, and there is no absolute agreement on how cell types should be formally and empirically defined (Various, 2017).

Globally, the increasing adoption of scRNA-seq is due to its multi-gene and unbiased profile. It allowed for the first time the non-directed profiling of molecules driving heterogeneity in cellular populations. Nonetheless, its use for defining cell identity still has some drawbacks. Even though the cost of high-throughput sequencing keeps dropping, single-cell RNA-seq still requires costly protocols, especially at the scale that it is currently performed for cell type discovery. This however can be mitigated by more targeted approaches, aimed at characterizing specific subsets of already known cell types isolated by their broad markers. scRNA-seq is also prone to batch effects, which can become more pronounced when comparing or integrating data generated by different protocols. This has been a very active topic of research,

and several batch alignment and correction methods can now account for these integration of different protocols (Butler et al., 2018; Haghverdi et al., 2018; Park et al., 2018; Stuart et al., 2019). From the protocol side, sample barcoding for multiplexed processing also greatly reduces batch issues (Shin et al., 2019; Stoeckius et al., 2018).

One last concern, although perhaps the largest, is the fact that profiling a tissue or a cell type with scRNA-seq does not inherently give any functional information about the cells. Cellular function has been from the beginning the major point to categorize cells. RNA, despite being easily correlated with protein presence, is not in most cases the effector molecule in a biological process. Additionally, most single-cell methodologies destroy the cell without imaging it, making the link between molecular makeup and morphology harder to obtain. While this is an ongoing research topic, profiling cells through the use of multi-omics technologies can help obtain a deeper mechanistic characterization. Information on open chromatin regions (Buenrostro et al., 2015), histone modifications (Kaya-Okur et al., 2019) or surface proteins (Stoeckius et al., 2017) have the potential to be combined, directly or indirectly, with single-cell RNA-seq (Clark et al., 2018). This can provide information on how these molecular layers interplay and learn about the intrinsic regulatory processes of gene expression (Gorin et al., 2019; Qiu et al., 2019). CRISPR screens with single-cell expression readout can also reveal more about cellular function (Datlinger et al., 2017; Dixit et al., 2016). Lastly, developments in spatial transcriptomics hold the promise of providing spatial context to cellular transcriptomes profiled individually, providing information on the tissue context for cell identity determination (Rodriques et al., 2019; Vickovic et al., 2019). Overall, while the discussion about where to draw the line between cell types still lasts, technological developments provide us with ever increasing information to approach a decisive and informative definition.

1.3 Methods for cell type classification

Single-cell RNA-seq was initially developed to obtain the whole transcriptome from samples with very low starting material (Tang et al., 2009). Nonetheless, the notion of using it to define cell types through their transcriptome was very early on envisioned. In 2011, Islam and colleagues end the discussion on their newly developed scRNA-seq method (STRT-seq) by stating "We envisage the future use of very large-scale single-cell transcriptional profiling to build a detailed map of naturally occurring cell types, which would give unprecedented access to the genetic machinery active in each type of cell at each stage of development." (Islam et al., 2011). The exponential increase

in the number of cells profiled per experiment eventually made this prediction come true. A large amount of single-cell projects have used the technology to profile cells captured from various tissues, in steady-state or disease conditions. Yet the most direct example of how this quote reflects the evolution of the field is the Human Cell Atlas (HCA) (Regev et al., 2017). This consortium has been established as a forum for scientists around the world to share their expertise on genomics, bioinformatics, and tissue biology, and coordinate the high-throughput profiling of cellular heterogeneity in the human body. The HCA has groups focusing not just on individual organs, but also on development (Behjati et al., 2018; Taylor et al., 2019) and disease.

In parallel, there have been increased efforts to obtain similar references for other species, in particular animal models (Cao et al., 2017; Fincher et al., 2018). The data collected for these species tends to have a greater cell coverage since the tissue samples can be more readily available. Furthermore, these atlases are by no means less important or useful than the human reference. The cell atlases produced for mouse (Han et al., 2018; Various, 2018) were of especial relevance, since they constitute the first broad, multi-organ cellular census of a mammalian organism, and one for which a large portion of biomedical science has relied on. The accessibility of human tissues for profiling and *in vitro* testing will be crucial in the near future. Nonetheless, having a mouse reference that can be related to human can not only teach us about the evolutionary principles that shape cell type evolution through gene expression, but also serve as a bridge to transpose mouse-based biomedical discoveries into a human context.

For a cell atlas to be used as a reference, it needs not only the expression data to be annotated, but also a computational framework that can use it to classify new datasets of interest. Over the last two years, several methods have been developed to handle scRNA-seq data (a comprehensive list can be found in Table 1.2), which can be added to other general purpose classification methods. These methods vary in complexity, but in general they rely on machine learning approaches to map the reference cell labels to the target dataset. While the most accurate method for this classification is still up for debate (see (Abdelaal et al., 2019; Köhler et al., 2019) in addition to benchmarks in individual method papers), there is agreement about the major challenges for this task. Classification methods should be aware of batch differences, be they caused by use of different scRNA-seq protocols or other technical differences in tissue processing. Different cell isolation and library preparation protocols can have a large impact on the number and type of genes detected (Mereu et al., 2019).

Table 1.2: Comprehensive list of papers detailing methods for automated cell state matching

Method Name	Short Description	Reference
scmap	k-nearest-neighbor search with cosine distance	(Kiselev et al., 2018)
matchSCore	Jaccard Index for cluster markers	(Mereu et al., 2018)
ClusterMap	Hierarchical clustering with marker gene binary expression	(Gao et al., 2018)
CaSTLe	XGBoost classification	(Lieberman et al., 2018)
Moana	Linear SVM on (sub)clusters	(Wagner and Yanai, 2018)
SAVER-X	Autoencoder	(Wang et al., 2018)
scQuery	Neural network classifier	(Alavi et al., 2018)
PopAlign	oNMF, Gaussian Mixture model and Jeffrey's divergence	(Chen et al., 2018)
scGen	VAE and linear classifier	(Lotfollahi et al., 2018)
scVI	VAE and hierarchical Bayesian model	(Lopez et al., 2018)
scPred	SVM in principal component space	(Alquicira-Hernández et al., 2018)
SingleCellNet	Random Forest on binary marker expression	(Tan and Cahan, 2018)
CellAssign	Multi-variable model with marker genes and hierarchical Bayesian framework	(Zhang et al., 2019a)
ACTINN	Neural network	(Ma and Pellegrini, 2019)
scID	Linear Discriminant Analysis with marker genes	(Boufea et al., 2019)
SingleR	Spearman correlation with training data	(Aran et al., 2019)
Garnett	Elastic net multinomial classifier using markers from hierarchical cell types	(Pliner et al., 2019)
SCINA	bimodal distribution of signature genes,	(Zhang et al., 2019b)
Cell BLAST	Adversarial Autoencoder and nearest neighbour search	(Cao et al., 2019b)
scMatch	Correlation with individual sample or average of references	(Hou et al., 2019)
SuperCT	Neural network with binary expression	(Xie et al., 2019)
CellO	Hierarchical binary classifiers	(Bernstein and Dewey, 2019)
scCoGAPS & projectR	NMF and projection in that latent space	(Stein-O'Brien et al., 2019)
SciBet	Entropy test and Bayesian comparison of multinomial distributions	(Li et al., 2019a)
Seurat "Anchors"	CCA, L2-normalisation and mutual nearest neighbours	(Stuart et al., 2019)
LIGER	integrative NMF and joint clustering	(Welch et al., 2019)
cellHarmony	Correlation with cluster centroids of mean marker gene expression	(DePasquale et al., 2019)
CHETA	Correlation with marker genes of hierarchical reference	(de Kanter et al., 2019)
scPopCorn	Co-membership Propensity Graph and (joint) k-partition	(Wang et al., 2019)
p-DCS	Voting based on known marker genes	(Domanskyi et al., 2019)
EnClaSC	Ensemble neural network classifier	(Chen et al., 2019)
scClassify	Ensemble classifier from inferred cell type tree	(Lin et al., 2019)

Many methods also mention the need to build a comprehensive reference, that should be integrated taking the into account the technical variability mentioned above. Training, and especially the prediction phases of the method should also be scalable. Models can take a very long time to train on larger references, and prediction steps that involve extensive manipulation or transformation of the target data can become time consuming with the ever growing size of expression matrices.

Lastly, some methods try to approach this classification problem from a hierarchical point of view (Lin et al., 2019; Pliner et al., 2019; Wagner and Yanai, 2018). This is based on the notion that cell types can be organised trees depicting phenotypic relationships. These trees represent not just developmentally-related lineages, but also the increasing specification of cellular function (still mostly correlating with terminal differentiation). This can be of great value in instances like describing cells from the immune system or the brain, where functional diversification leads to more intricate phenotypes (see Section 1.4). Notwithstanding, a hierarchical classification can also be seen as a method that reflects the uncertainty in the prediction. Each individual cell ideally conforms to a determined phenotype, which would correspond to a leaf node in an ideal cell hierarchy. Assigning a cell to a parent node rather than a terminal one (or not doing it with a high confidence) can be caused by data sparsity or low coverage, and thus not necessarily reflecting a naturally occurring hierarchy of gene expression-driven cellular phenotypes. Yet this structure is intuitive and informative, and projects like the Cell Ontology have considerable value in creating a controlled vocabulary to name and relate cell types (Bard et al., 2005), with some of the methods listed here explicitly conforming to it. The use of a curated and specific nomenclature should thus be incentivized when doing *de novo* annotation of scRNA-seq data, and supplying these labels can greatly accelerate the data interpretation and its application in the development of new algorithms.

Large collections of data and development of informative references can be of use in multiple ways. A steady-state cell identity reference can serve as a baseline to which a disease sample can be compared. Having a sufficiently comprehensive cell registry can do away with the need to generate a reference dataset if the goal is quantifying alterations to the proportions of known cell populations. Evolutionary biology can also benefit from predictive models for cell identity. Models can be adapted to function across species, which can help trace the evolutionary origins of cell types. Producing interpretable models from integrated data can also be informative in itself. Some models return the importance of genes or gene sets in classifying each cell type, and as such can help uncover novel features of a cell's phenotype. Finally, organised

references can also speed up new in-depth studies of specific cell types, as well as studies focusing on other aspects of cell identity (e.g. open chromatin, methylation, proteome, or spatial interactions). It should be noted that the methods discussed so far in this section, while being in their majority developed for scRNA-seq, can also for the most part be adapted to other data modalities like scATAC-seq (for open chromatin) or CITE-seq (combining RNA and surface protein detection). Modelling cell identity with multiple layers can revel more details about the molecules shaping it, how they interact, and their relative importance.

1.4 Cell identity in the immune system

The immune system is one of the most complex and diverse biological systems across the animal kingdom. The increased evolutionary pressure caused by the need to continuously adapt to the fast evolving pathogens (Barreiro and Quintana-Murci, 2010) has resulted in a broad variety of molecular pathways and cells. The variability in the types of cells found in the immune system is directly related to their intrinsic plasticity in gene expression. Immune cells are very responsive to their environment, having to constantly fine-tune expression programmes to react in a prompt and targeted manner. It then comes as no surprise that many cell states have been determined and named in immunology, and it is, perhaps on par with neurobiology, the field where the definition of cell type and cell state clash the most.

Due to the fact that immune cells are non-adherent cells, immunology benefited immensely from the development of flow cytometry. Immune cells have been deeply characterised by this technology, with antibodies targeting surface receptors as well as cytoplasmic proteins. It then comes as no surprise that the immune system has been an early and major target of single-cell sequencing methods. scRNA-seq has had a role in the fine-grained mapping of gene expression changes in haematopoiesis (Watcham et al., 2019), discovering and reorganising subpopulations (Villani et al., 2017), mapping their heterogeneity across tissues (Miragaia et al., 2019; Scott et al., 2018), studying immune response to pathogens (Lönnberg et al., 2017; Stubbington et al., 2016), and map communication of immune cells with their tissue of residence (Vento-Tormo et al., 2018).

Immunity can be divided into innate and adaptive. The latter depends on a subset of lymphocytes which are responsible for an immune response that can flexibly adjust to invading pathogens in a non-evolutionary way (i.e. without the need for selection at the level of the individual). The key strength of this system is the use of receptors

which recombine and mutate (Krangel, 2009), forming a highly diverse repertoire that can eventually be selected to respond to particular invaders. This variability, central to the adaptive immune response, is further complemented by immune memory, that is, the specific repertoire obtained when combating an infection will remain stored in the organism in the form of inactive immune cells, which can be more quickly reactivated should the same threat reappear. This is far more advantageous than having to undergo selection of the receptor repertoire every time the same pathogen is introduced in the system.

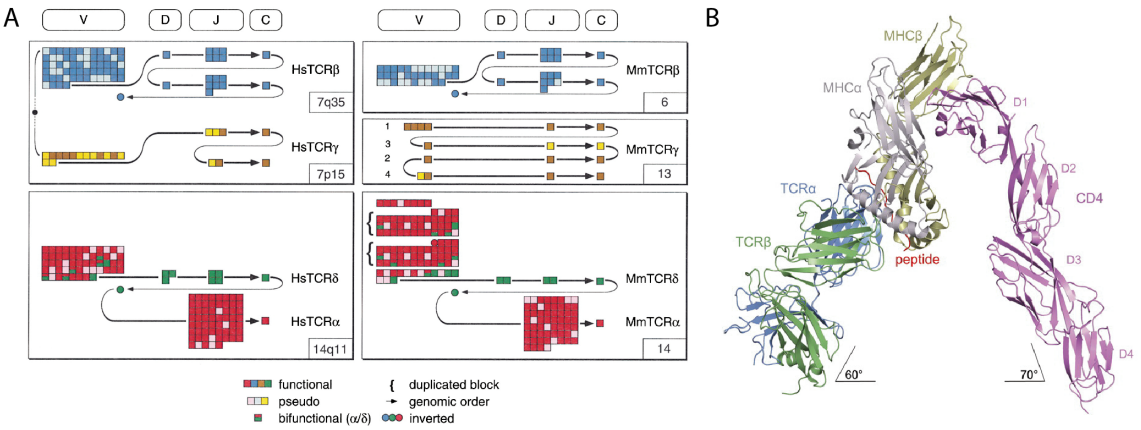


Fig. 1.2: **Gene and protein structure of TCR**
(A) The genomic organization of the human (left) and mouse (right) TCR genes α (red), β (blue), γ (brown), and δ (green), showing clusters of V, D, J, and C gene segments aligned vertically for clarity. Arrows represent the direction of transcription within each of the TCR genes; squares and circles indicate gene elements in the direct and reverse orientations, respectively. The murine TCR γ2 gene is inverted relative to the rest of the locus. Dark colors indicate apparently functional gene elements, while lighter shades represent pseudogenes. Curly brackets indicate the duplicated sets of V genes in murine TCR α/δ locus. The TCR β and TCR γ loci are both on human chromosome 7, on opposite sides of the centromere (schematically represented by the black circle). *Original figure published in (Glusman et al., 2001).*
(B) Ribbon diagram of the complex oriented as if the TCR MS2-3C8 and CD4 molecules are attached to the T cell at the bottom and the HLA-DR4 MHC class II molecule is attached to an opposing APC at the top. TCR α chain, blue; TCR β chain, green; CD4, pink; MHC α chain, gray; MHC β chain, yellow; MBP peptide, red. *Original figure published in (Yin et al., 2012).*

Within adaptive immunity lymphocytes, T cells fill various niches, but are broadly considered to be the orchestrators of immune response (Kumar et al., 2018). T cells are characterised by their expression of the T Cell Receptor (TCR), a dimeric surface protein that can recognise an antigen presented by an Antigen Presenting

Cell (APC) (Reinherz, 2014). This receptor's ability to recognize a trove of antigens resides in the original gene's unique recombination capacity. TCR genomic segments are composed of many genes (in addition to a constant region) - grouped into variable (V), diversity (D) and junction (J) genes - that encode the variable section of the final protein, which interacts with the antigen presented by the MHC complex (Figure 1.2A). During T cell development in the thymus, these genes are recombined through the action of RAG enzymes, which target recombination signal sequences to cleave DNA and join them - first D and J (if D is present), then (D)J and V. The insertion of additional non-templated nucleotides at the junctions can result in further variability. There are numerous V and J genes, which gives rise to a large number of possible V-J combinations, thus ensuring the diversity needed for antigen recognition by T cells. This is further augmented by differential combination of TCR chains in the final receptor. The activity of each receptor sub-unit is subject to selective pressures that ensure that it can functionally recognise and respond to foreign antigens, while being unresponsive to self-produced peptides and thus avoid auto-immune responses. In adaptive T cells, these receptors are composed of an α and a β chain. γ and δ chains also exist as a pair, but are less variable which results in a different type of response (Simoes et al., 2018).

The TCR is part of a larger membrane surface complex that assists in the recognition of the antigen being presented, as well as the APC presenting them (Figure 1.1B). T lymphocytes can thus be separated into two subsets with a shared developmental origin, bifurcating depending on the type of antigen-presenting Major Histocompatibility Complex (MHC) they can match. Consequently, each with their own APC matching capabilities and is easily identifiable by the expression of a surface protein that participates in this specific interaction. CD8-expressing T cells recognise antigens presented through MHC class I, which exists on the surface of almost all cells. This recognition elicits the maturation of $CD8^+$ T cells, preparing them for an anti-cellular response. This subset is accordingly also named cytotoxic, and through the use of perforins and granzymes they destroy cancer cells, as well as cells infected by intracellular pathogens (Halle et al., 2017).

$CD4^+$ T cells are the other lineage of T cells. Also known as T-helper (Th) cells, these lymphocytes are credited with the organisation of immune response, producing cytokines that serve as triggers or blockers of particular immune reactions (Luckheeram et al., 2012). Th cells recognise antigens presented by the MHC class II, present only on the membrane of dendritic cells, mononuclear phagocytes, some endothelial cells, thymic epithelial cells (important during T cell selection for functional,

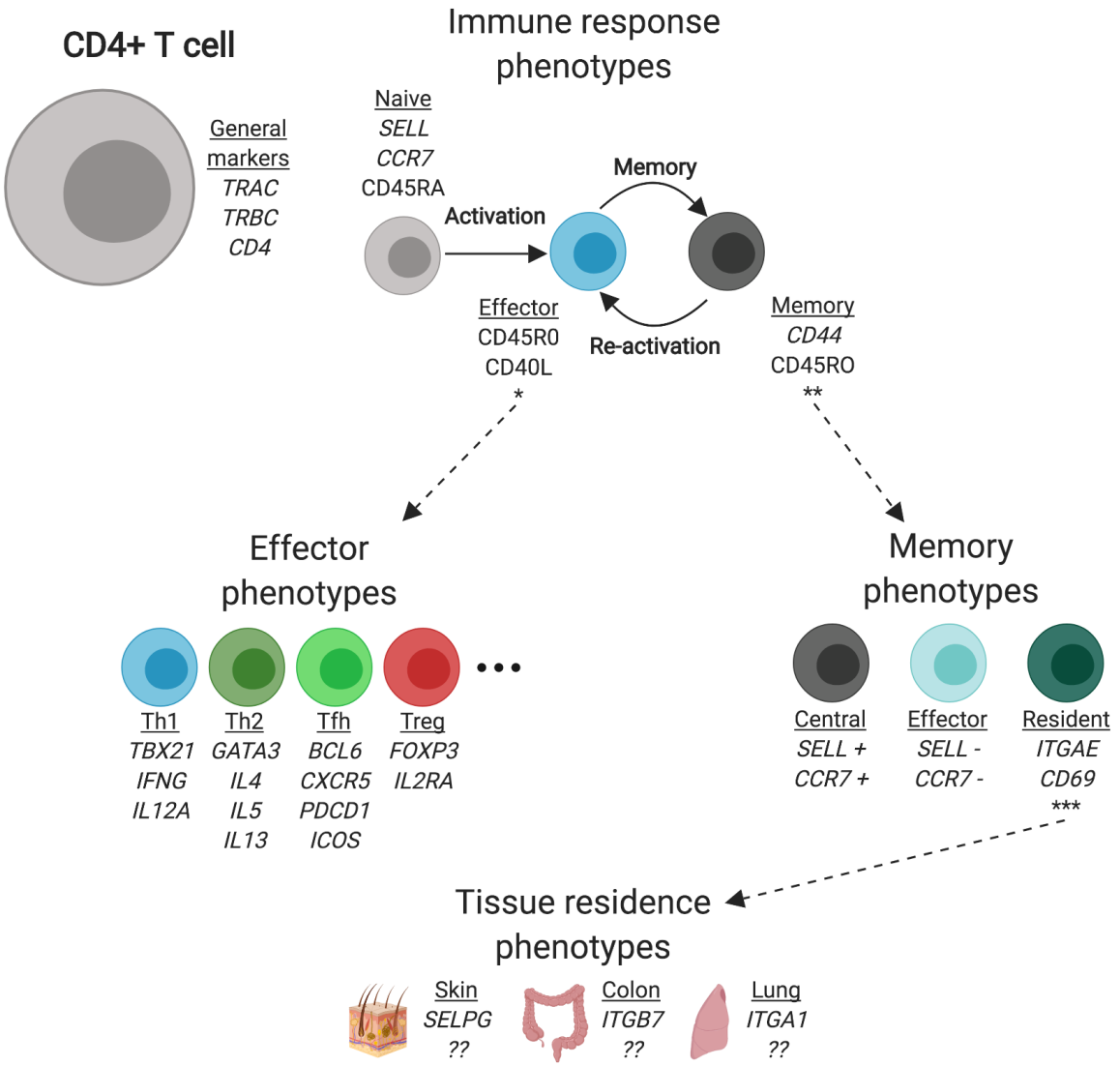


Fig. 1.3: An overview of known T-helper cell heterogeneity and key marker genes. Beyond their core markers, Th cells can be classified into based on different phenotypes that depend on stage of immune response, the type of effector function, the type of memory cell they form and their tissue of residence (a topic understudied comparatively to the rest). Question marks (??) represent unresolved phenotypes.

non-self-responding TCR), and B cells. This interaction, combined with signalling from the media where the cell is acting, induce an activation programme of the cell that is specific to the external threat being handled. $CD4^+$ T cells encompass a large transcriptional plasticity, which results in diverse related phenotypes (Figure 1.3). Th cells have classically been organised into various effector phenotypes based on

their cytokine secretion profile (Mosmann et al., 1986; Schmitt and Ueno, 2015), allowing them to orchestrate the professional immune cells in the microenvironment. For instance, IFNγ production by Th1 cells has been identified as a key signalling molecule to combat intracellular parasites through the stimulation of macrophages, as well as class switch recombination of B cells to an IgG isotype. In turn, Th2 cells use IL-4 and IL-13 to stimulate basophils and mast cells to release granules against helminth invaders, and Th17 cells coordinate neutrophil recruitment by epithelial cells through IL-17A and IL-17F (Weaver et al., 2013). While diverse in function, these effector phenotypes are not the sole drivers of variability between Th cells, which also vary according to their activation state (naïve, effector, and memory cells) and with the host environment cues (tissue-specific phenotypes).

Upon finishing responding to an infection, T cells can go into a lowly replicative memory state in which they will save the TCR that drove the specialized response. The various memory states relate to the level of activation of the cell, but also to its tissue of residence. Cells expressing the chemokine receptor CCR7 are in a more naive, non-stimulated state, and also target lymphoid tissues like lymph nodes or the spleen, where most of antigen presenting to $CD4^+$ T cells takes place. In addition, tissue-homing and residency phenotypes exist, all of them characterised by the involvement of one or more chemokine receptors or adhesion molecules like integrins. Nonetheless, tissue-specificity in T-helper cells, and even more broadly in immune cells, is still generally understudied. Recent developments using single-cell high throughout methods have tackled this questions (Scott et al., 2018; Wong et al., 2016a), and it is expected that future efforts will rely on the accumulation of data to extract these patterns from cross-tissue samples.

Among the phenotypic variability of T-helper cells we can find the particular subset termed T-regulatory (Treg) cells. They are different from most Th cell subtypes in that, rather than boosting immune response, they are responsible for dampening it (Sakaguchi et al., 1995). This regulatory role in the immune system is of dire importance. Leaving the immune response unchecked can lead to destructive responses that will adversely affect the organism, as in autoimmune diseases. Treg cells were originally identified by their high expression of CD25, but as a subset they are more clearly defined by the expression of the FOXP3 transcription factor (Hori et al., 2003). Despite the focus on $CD4^+$ Treg cells here presented, $CD8^+$ cells can also have a regulatory phenotype, yet this are understudied compared to its $CD4^+$ counterpart (Yu et al., 2018).

Further subsets of Treg cells have been described, either related to the various parallel programmes that Th cells can adopt or their developmental origin. All T cells derive from a Common Lymphoid Progenitor cell that originates through haematopoiesis in the bone marrow, and travels via the bloodstream to mature in the thymus, where their TCR recombines and is tested for responsiveness to foreign antigens (positive selection) and against self-antigens (negative selection). However, natural Treg cells are derived from a subset of T cells with an intermediate level of response to self-antigens. This subset is further supplemented by induced Treg cells, which originate from other T-helper cells. While both natural and induced T-regulatory cells share a role, their distinct origins extend their TCR repertoire and thus their function (Zhang et al., 2014). Beyond this, Treg cells are also subject to memory and tissue-trafficking phenotypes like the remaining Th cells (Huehn et al., 2004), although these are not as well studied.

Immune cells are also described to have roles beyond defense against pathogens. These roles involve interactions with other non-immune tissues and mostly focus on their maintenance (Gordon and Martinez-Pomares, 2017; Laurent et al., 2017), and the immune system has also been described as relaying signals to the nervous system (Veiga-Fernandes and Mucida, 2016). Treg cells have been increasingly noted to be relevant, not just for their role in the immune system, but also for their functions beyond it. This regulatory subset has been shown to be involved in tissue repair (Li et al., 2018b) (chiefly muscle (Burzyn et al., 2013)), hair growth (Ali et al., 2017), and homeostatic regulation of gut microbiota (Cebula et al., 2013) and adipose tissue (Cipolletta, 2014; Sharma and Rudra, 2018). These functions, being widespread in the organism, consequently rely on an efficient trafficking and tissue localization scheme (Liston and Gray, 2014). Despite the importance of understanding how these migration and adaptation programmes are constituted and regulated (Agace, 2006), this aspect of the immune system is still incompletely understood.

1.5 Tissue-specific gene expression

Histological studies have uncovered many details of organ biology and physiology. Tissue staining is routinely used in pathology, and a better understanding of which molecules are markers of different tissue structures and cells in steady-state has resulted in important medical advancements.

Early studies in transcriptomics using microarrays dissected transcriptional responses to metabolic shifts (DeRisi et al., 1997) and disease (with a particular focus in cancer) (Rhodes et al., 2004), with homeostatic tissue sample comparison only appearing later (Shyamsundar et al., 2005).

RNA-sequencing has, from its inception, been linked to the unraveling of cross-organ and tissue differences (Mortazavi et al., 2008). Compared with preceding technologies, RNA-seq was capable of detecting a broader variety of transcripts in an unbiased way, along with high confidence splice junctions and allele-specific expression, with the added benefit of doing it for a lower cost (Wang et al., 2009). RNA-seq was quickly adopted and improved (see Section 1.2), extending its sensitivity and breadth of applications. Consortia were developed around the use of sequencing technologies for different biomedical purposes, often with RNA-seq taking a pivotal role (Lonsdale et al., 2013; The Cancer Genome Atlas Research Network et al., 2013; The ENCODE Project Consortium, 2012). These large collections of data were instrumental in revealing the functionality of genomic regions and relationships between samples. With data from the Genotype-Tissue Expression (GTEx) consortium, it was revealed how human tissues transcriptionally relate to each other, as well as what genes vary in expression across tissues and individuals (Melé et al., 2015). The Cancer Genome Atlas (TCGA) relied on RNA-seq, as well as other data modalities, from several cancer types to map the similarities between different tumours, and identify potentially important pathways for the treatment of those malignancies (Hoadley et al., 2018). Comparison between disease samples and steady-state can also be particularly informative, for example in understanding how tumours affect their adjacent tissue (Aran et al., 2017), or how tumour growth compares to developmental tissues and which pathways are involved (Young et al., 2018). In short, while large databases of expression data can serve as useful resources for broader applications by the scientific community, they can also be mined for emerging patterns.

Transcriptomic data can also be analysed beyond one species to gain understanding of the evolutionary links of gene expression programmes. Early microarray data analysis showed how human-chimpanzee divergence was especially accentuated when looking at brain RNA (Enard et al., 2002). Collection of samples from more species, combined with the use of RNA-seq, augmented the resolution of what gene expression changes could be observed (Brawand et al., 2011). Varying divergence rates for different tissues, gene groups and genomic regions, could be observed and associated to different selective pressures and tissue functions. Further studies have since compared other species (Li et al., 2014) or aspects of the transcrip-

tome (Barbosa-Morais et al., 2012), revealing the intricate way evolution sculpted molecular programmes in different tissues across the tree of life, and defined the core genes involved in tissue function.

The functional associations observed between tissues are a consequence of the similarities and differences of the cell types that constitute them. These are mostly a result of the developmental processes giving rise to these tissues. For example, most tissues contain epithelial cells, marked by EPCAM, which share certain features such as forming barriers and secretory functions (Trzpis et al., 2007). Epithelial cells have been found to be vastly diverse within and between tissues, adopting different shapes and spatial arrangements (Wang et al., 2012), as well as further cytological changes adapted to the specific tissue biology.

Many aspects of tissue-specific heterogeneity stem from immune cells, perhaps owing to their mobility and plasticity. Various tissue-specific functions of Treg cells have been described above (Section 1.4). Macrophage heterogeneity represents another paradigmatic case of between-tissue phenotypic variability. In adult humans, circulating macrophages derive from bone marrow progenitors; in contrast, tissue-resident macrophages have been demonstrated to be developmentally related to haematopoietic progenitors in the yolk sac (Gomez Perdiguero et al., 2015). These macrophage subsets are important in mediating tissue immunity, while in parallel governing their homeostasis, such as synaptic pruning by microglia, heme recycling by splenic macrophages, or the pro-angiogenic role of Hofbauer cells at the maternal-fetal interface. Importantly, tissue-specific functions are a consequence of signalling in the local environment, which is capable of completely reprogramming macrophage chromatin, gene expression and function (Gosselin et al., 2014; Lavin et al., 2014), and consequently influence their response to tissue-specific injuries (Hoyer et al., 2019). This heterogeneity has also been detected within tissues, and in the gut has been associated with signalling provided by local neurons (Gabanyi et al., 2016). Single-cell RNA-sequencing has also been used to reveal cross-tissue conserved regulators of macrophage identity (Scott et al., 2018), and could in the future be used to further explore potential subpopulation heterogeneity and correlate it with gene expression spatial data to identify associations with specific anatomic locations within organs.

The application of scRNA-seq methods can extend these methods to comparisons between cell types, which results in larger scale comparisons, yet will open a window into how different programmes are specified for cell function in evolution and how they translate across species. It has recently been showed how variability in expression

relates to evolution of innate immune response in fibroblasts (Hagai et al., 2018). Data from this study has been further used to test an artificial intelligence method that was capable of accurately predict species-specific responses solely based on the data from the remaining organisms sampled (Lotfollahi et al., 2018). As well as understanding evolutionary biology of cell types or immune responses, these types of studies and applications can have considerable impact in translating results from model organisms into the clinic.

1.6 Insights and scope of this thesis

Single-cell RNA-seq has revolutionized the profiling of cell type heterogeneity over the last decade. This has allowed for a deep, unbiased look into several organs and organisms, profiling hundreds of cell types at higher resolution. At the same time, progress has been made in computationally combining datasets for further analysis. As an increasing number of scRNA-seq datasets is produced, we come ever closer to a first draft of a transcriptional Human Cell Atlas, showcasing the full spectrum of cellular variety in our species.

The expansion in cell throughput is now permitting the study of smaller, rarer subpopulations. While specific cell types can still be sorted prior to sequencing for deeper profiling, unknown and underrepresented cell types will require larger numbers to be detected. This profound transcriptional portrayal of cells also often results in valuable resources that can be examined for functional targets of novel therapies and assays, which is especially true when studying immune cells. Developing directed cell therapies is a long-term goal of many medical fields, but a thorough knowledge of key cell types is still needed.

A transcriptional reference for cell types can be a key resource for those employing scRNA-seq. Having a ready-to-use resource that draws on the combined knowledge of the data generated would provide immediate assistance for automatic annotation of novel projects. Additionally, an exhaustive and integrated collection can be very informative about cell and tissue biology. However, the limits of this integration should also be tested and examined.

After this introductory chapter, Chapter 2 will show a deep dive into T-regulatory cell heterogeneity using single-cell RNA-seq. Treg cells have been shown to have critical roles in steady-state and disease, but it is still not fully understood which subpopulations fulfill which functions in different tissues, and how this heterogeneity relates to cross-tissue diversity. The chapter will describe Treg cell subpopulations

detected in mouse in different tissues and how they compare to other resident T-helper cells. These subpopulations reflect different activation states, and form a phenotypic continuum between peripheral tissues (skin and colon) and their respective draining lymph nodes. The first sections will also discuss the limits of heterogeneity detection using scRNA-seq, especially when using two different protocols. Lastly, a mouse-to-human comparison will be presented, comparing conservation and divergence of gene programmes and Treg cell subpopulations.

Chapters 3 and 4 will focus on the use of broad scRNA-seq data collections to create informative references for automatic cell type annotation. Chapter 3 will detail the development of *CellTypist*, a pipeline to integrate diverse scRNA-seq datasets and cluster them into meaningful groups that approximate commonly defined cell identity, and the training of an updatable classifier that can be used to annotate new datasets. All annotation data available from these datasets is also collected, and the classifier train is also in itself informative. Following this, Chapter 4 will be centred on the dissection of a large collection of human scRNA-seq data. After application of *CellTypist*, it will explore how gene expression at the cell type level influences tissue similarity, as well as uncover the groups of genes characterising cell identity.

This thesis ends in Chapter 5, where I will be discussing the broader picture of the results reported in this thesis. This chapter will explore to what detail cell identity can be deconstructed, and what that means for informative automated annotation of new datasets, as well as to our understanding of cell biology and how they are categorized.

Chapter 2

Tissue adaptation of T-regulatory cells

Non-lymphoid tissues (NLTs) harbour a pool of adaptive immune cells with largely unexplored phenotype and development. We used single-cell RNA-seq to characterise 35000 $CD4^+$ regulatory (Treg) and memory (Tmem) T cells in mouse skin and colon, their respective draining lymph nodes (LNs) and spleen. In these tissues, we identified Treg cell subpopulations with distinct degrees of NLT phenotype. Subpopulation pseudotime ordering and gene kinetics were consistent in recruitment to skin and colon, yet the initial NLT-priming in LNs and the final stages of NLT functional adaptation reflected tissue-specific differences. Predicted kinetics were recapitulated using an *in vivo* melanoma-induction model, validating key regulators and receptors. Finally, we profiled human blood and NLT Treg and Tmem cells, and identified cross-mammalian conserved tissue signatures. In summary, we describe the relationship between Treg cell heterogeneity and recruitment to NLTs through the combined use of computational prediction and *in vivo* validation.

This chapter has been published in *Immunity* as *Single-cell transcriptomics of regulatory T cells reveals trajectories of tissue adaptation* (Miragaia et al., 2019), with the exception of Section 2.2.6 and parts of Sections 2.3 and 2.4. The Methods section in this chapter only includes the computational steps. The remaining experimental methods, as well as the supplementary figures, can be found in Appendix A.

Additional contributions: experiments in this chapter were performed by Ricardo J Miragaia. The study was designed by Ricardo J Miragaia, Sarah A Teichmann, Agnieszka Chomka, Fiona Powrie, and myself. Ricardo J Miragaia is a leading co-author of the manuscript. Full acknowledgements can be found in Appendix A.

2.1 Introduction

T-regulatory (Treg) cells are a specialised $CD4^+$ T cell subset which control immune responses and play a central role in homeostasis (Izcue et al., 2009; Sakaguchi, 2004). Recent studies have described unique tissue-specific adaptations of non-lymphoid tissue (NLTs) Treg cells distinct from their lymphoid tissue (LT) counterparts. This includes acquisition of an effector phenotype with expression of transcripts encoding effector molecules (*Ctla4*, *Gzmb*, *Klrg1*), chemokines and their receptors (*Ccr4*), and immunosuppressive cytokines (*Il10*) (Bollrath and Powrie, 2013; Panduro et al., 2016), in addition to tissue-specific signature genes associated with their role in each environment (Liston and Gray, 2014). Nonetheless, their full transcriptional phenotype and its reflection on NLT population heterogeneity is yet to be uncovered.

Trafficking of T cells to NLTs occurs in steady-state conditions and development (Kimpton et al., 1995; Thome et al., 2015) as well as in response to harmless stimuli at barrier surfaces such as commensal bacteria and dietary antigens (Ivanov et al., 2008). Treg cell migration requires tissue-specific cues involving integrins, chemokine and other G-protein coupled receptors (Cepek et al., 1994; Chow et al., 2015; Kim et al., 2013).

To provide a deeper insight into Treg cell populations in NLTs, we analysed single-cell RNA-seq (scRNA-seq) data of Treg cells from mouse colon and skin, and compared them to LT populations. We identified various transcriptionally distinct clusters of Treg cells in LTs and NLTs, namely a subpopulation in the LTs which showed heavy priming to the NLT environment. Pseudotime ordering of these subpopulations further revealed the transcriptomic adaptations occurring in Treg cells during their transition from the lymph node to barrier tissues. Our results show that these steady-state adaptations share a core signature between bLN-to-skin and mLN-to-colon trajectories, indicative of a general NLT residency programme in barrier tissues. These findings were recapitulated during *de novo* Treg cell recruitment to melanoma in a murine model system. Lastly, we examined the evolutionarily conservation of NLT Treg cells' identity between mouse and human.

2.2 Results

2.2.1 Treg and Tmem cell identity in NLTs is driven by a common expression module

We performed scRNA-seq on isolated $CD4^+Foxp3^+$ (Treg) and $CD4^+Foxp3^-CD44high$ memory (Tmem) T cells (Figure A.1A) from two barrier NLT sites - the colonic lamina propria (hereinafter referred to as colon) and the skin - their lymphoid counterparts in the draining mesenteric and brachial lymph nodes (mLN and bLN), and the spleen from a Foxp3-GFP mouse reporter line (Bettelli et al., 2006) (Figure 2.1A). We will refer to Treg and Tmem cells together as $CD4^+$ T cells. For each sorted population, single-cells were captured using the droplet-based microfluidic system Chromium (10x Genomics), hereinafter referred to as 10x. We obtained 30396 good quality cells (see Methods, Figure A.1C, Table A.5). Using the same gating strategy, two Smart-seq2 (Picelli et al., 2014) plate-based datasets were produced independently. These confirmed findings drawn from the 10x, and complemented them with higher gene coverage and full T cell receptor (TCR) sequences.

A tSNE projection (Figure 2.1B) after filtering (Figure A.1B) showed a division between LT and NLT, with cells from LTs divided into two clusters, according to cell-type. NLT cells formed one single skin cluster and two clusters separating Treg and Tmem cells from colon (Figure 2.1B). We defined gene expression signatures for Treg and Tmem cells in peripheral tissues by examining differentially expressed (DE) genes between all NLT and LT cells and, in parallel, between Treg and Tmem cells (Figure 2.1C). NLT T cell populations are characterised by the expression of several elements of the TNFRSF-NF-κB pathway, including transducers (*Traf1*, *Traf4*, *Traf2b*), effectors (*Nfkb1*, *Nfkb2*, *Rel*, *Rela*, *Relb*) and inhibitors (*Nfkbib*, *Nfkbid*, *Nfkbie*). In Tmem cells, these were accompanied by cytokines (*Tnfsf8*, *Tnfsf11*) and various pathway inhibitors, such as *Tnfaip8*. In contrast, NLT Treg cells expressed TNF receptors (*Tnfrsf4*, *Tnfrsf9*, *Tnfrsf18*) and transducers (*Pim1*), underscoring the importance of signalling via the TNFRSF-NF-κB axis in controlling Treg cells in the peripheral tissues. Several chemokine receptors appeared DE across tissues and cell types. *Ccr4*, *Ccr8* and *Cxcr4* were upregulated in both colon and skin T cells, while *Ccr1* and *Ccr5* were specific to colon and *Ccr6* to skin. *Cxcr6* was more highly expressed in NLT Tmem cells. We also detected other genes involved in NLT identity (*Crem*, *Rgs2*, *Il1r2*, *Icos*, *Hif1a*, *Kdm6b*, *Gata3*), including some specific to Tmem (*Vps37b*, *Id2*, *Ramp3*, *Tnfsf8*) and Treg cells (*Il10*, *Gzmb*, *Ctla4*, *Cd83*, *Socs2*).

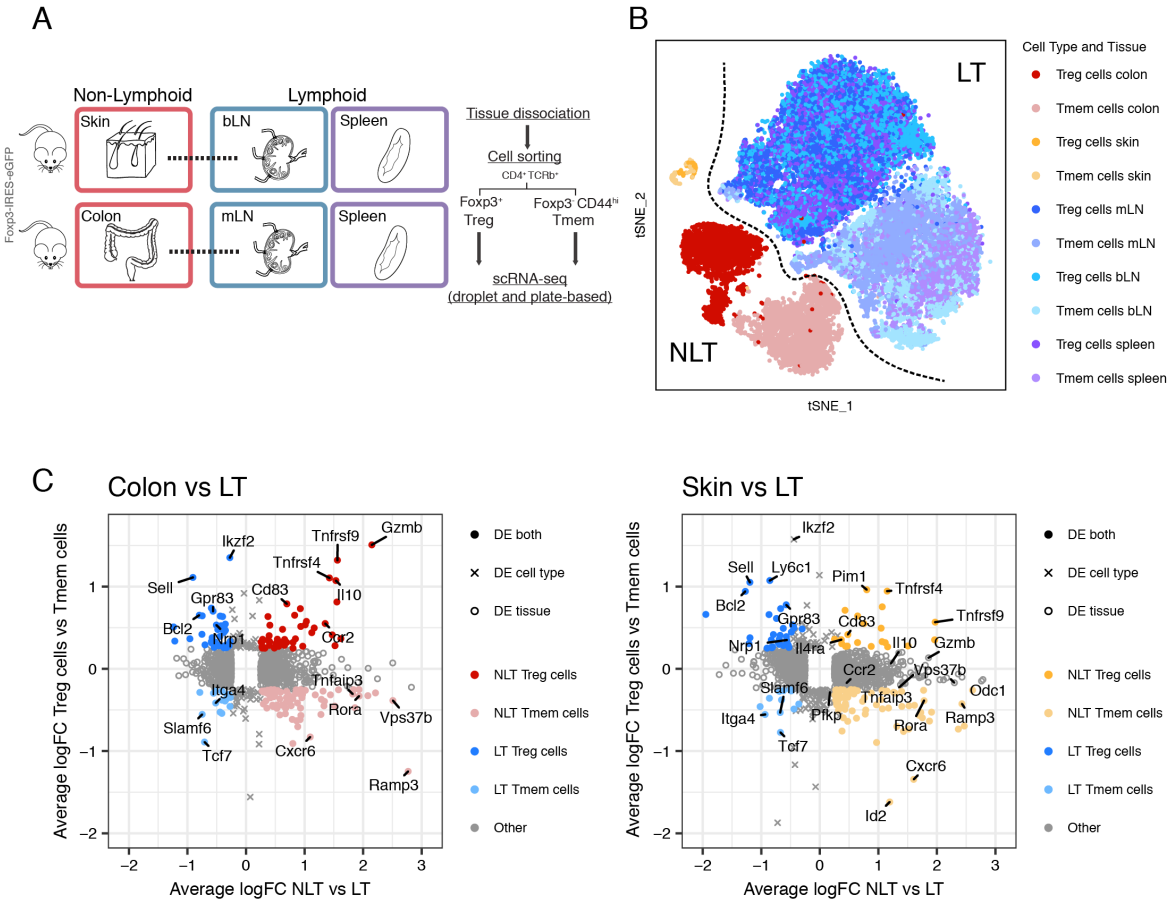


Fig. 2.1: **Steady-state scRNA-seq datasets of CD4$^+$ T cells from LT and NLT** **(A)** Experimental design for scRNA-seq data collection. **(B)** t-SNE representing all Treg and Tmem cells that passed quality control. **(C)** Genes defining the identity of Treg and Tmem cells in lymphoid and non-lymphoid tissues. Colon and skin were individually compared with their corresponding draining lymph node and spleen cells. See also Figure A.1.

Together, the scRNA-seq datasets collected provide a comprehensive overview of Treg and Tmem cells in multiple lymphoid and non-lymphoid tissues, and identify the TNFRSF-NF-κB pathway as key to their barrier tissue identity.

2.2.2 Heterogeneity within LT and NLT Treg cell populations

Treg cell phenotypical and functional heterogeneity has been extensively discussed in recent years (Campbell and Koch, 2011; Josefowicz et al., 2012). Clustering our data within each tissue grouped Treg cells into distinct subpopulations (Figure 2.2A) with clearly defined marker genes (Figure 2.2B). Across lymphoid organs, we identified

central and effector Treg (cTreg and eTreg) cell subsets (Cretney et al., 2011; Vasanthakumar et al., 2015). cTreg cells express typical LT-associated markers, such as *Tcf7*, *Bcl2*, *Sell*, *S1pr1*, while eTreg cells expressed a subset of NLT-associated genes, like *Tnfrsf9*, *Relb*, *Ikzf2* and *Pdcd1*. We also detected a subpopulation of Treg cells with high expression of *Stat1* and interferon stimulated genes exclusively in the bLN. A fourth, less frequent population in lymphoid tissues (5-10%; Figure 2.2C), which we named Treg NLT-like cells, expresses eTreg cell markers as well as genes characteristic of NLT T cells, such as *Itgae*, *Rora*, *Fgl2*, *Klrg1* (Figure 2.2B). We hypothesize that this population is primed to migrate and adapt to NLTs. Indeed, DE genes between NLT-like Treg cells from mLN and bLN revealed that the colon-homing molecules *Ccr9* and *Itga4*, as well as their regulator *Batf* were upregulated specifically in the mLN, while *Cxcr3* and *Itgb1* were present in the bLN (Figure 2.2E). These differences were not observed between other LN subpopulations (data not shown).

To quantify the bias towards LT or NLT phenotypes, we calculated an NLT-LT marker gene signature for each cluster (Figure 2.2D; see Methods). Consistently across all LTs, cTreg cells exhibited a clear LT signature, while eTregs and NLT-like Tregs leaned towards an NLT profile, which was more pronounced in the latter.

In the colon, we found three subpopulations of Treg cells that we labeled as NLT, suppressive and LT-like. Treg NLT and suppressive cells were present in equal proportions, both exhibiting NLT traits (Figure 2.2C,D). Treg NLT cells in colon express higher amounts of *Gata3*, *Nrp1*, *Areg*, *Il1rl1*, *Ikzf2*, matching the known thymic-derived GATA3$^+$-subpopulation (Hu and Zhao, 2015; Schiering et al., 2014), while suppressive colonic Treg cells expressed more *Il10*, *Gzmb*, *Lag3*, *Cxcr3*, resembling the peripherally-derived RORγt$^+$-subpopulation (Ohnmacht et al., 2015; Schiering et al., 2014; Sefik et al., 2015). *Rorc* itself, while not present as a marker, appears in a higher percentage of Treg suppressive cells (6.16% vs 2.85% in colonic Treg NLT cells). Technical limitations for detection of lowly expressed genes by scRNA-seq might account for the difficulty in capturing *Rorc* transcripts. Lastly, LT-like Treg cells differed from other colonic populations by expressing LT-associated genes including *Sell*, *Ccr7*, *Tcf7*, *Bcl2*, and lower amounts of NLT-associated genes such as *Klrg1*, *Cd44*, *Icos*, *Rora*, *Tnfrsf9*, *Itgae* (Figure 2.2B).

In contrast to the colon, and likely as a consequence of fewer cells captured, skin Treg cells did not show evident heterogeneity (Figure 2.2A). They expressed an unequivocal NLT signature (Figure 2.2D), but it was not clear to which colonic Treg cell populations they were most similar (Figure 2.2B). We addressed this by using a logistic regression model to calculate the probability of each skin Treg cell

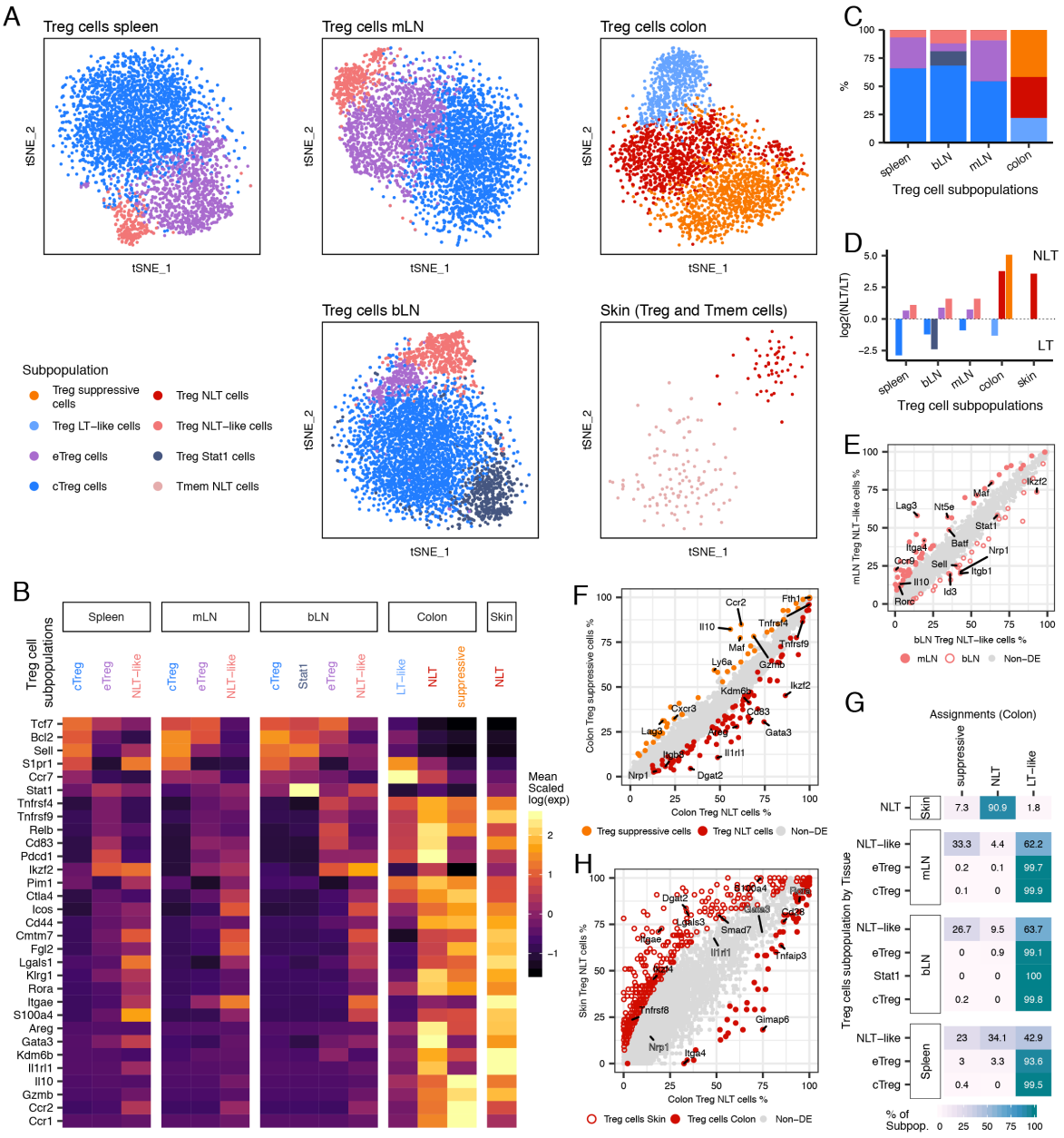


Fig. 2.2: **Heterogeneity within LT and NLT Treg populations**
(A) t-SNE projections of Treg cells per tissue, coloured by subpopulation. cTreg: central Treg, eTreg: effector Treg. **(B)** Subpopulation marker gene mean expression (z-score). Values greater than 2.5 or lower than -1.5 are coloured equally. **(C)** Relative proportions of Treg cell subpopulations within each tissue that revealed heterogeneity. **(D)** NLT/LT signature score in each Treg cell subpopulation, measured as the ratio between the number of NLT and LT genes that have been identified as significantly upregulated in each cluster. (Continued on the following page.)

Fig. 2.2: (continued) **(E)** Percentage of cells expressing each gene in Treg NLT-like cells from mLN and bLN. Genes that are upregulated in the bLN subpopulation are represented by an open circle, and genes upregulated in mLN are represented by a filled circle. **(F)** Percentage of cells expressing each gene in colon Treg suppressive and Treg NLT subpopulations. **(G)** Matching of non-colonic Treg cells to colonic Treg cell subpopulations using a logistic regression model (90% accuracy, see Methods). Table shows the percentage of each identified subpopulation (y-axis) that were labelled by the model as each Treg cell cluster (x-axis). **(H)** Percentage of cells expressing each gene in skin Treg NLT and colon Treg NLT cell subpopulations. See also Figure A.2.

identifying as one of the colonic subpopulations (Figure 2.2G, see Methods). This revealed that most skin Treg cells were more similar to colonic Treg NLT than to Treg suppressive cells. Accordingly, colon Treg NLT cell marker genes *Gata3*, *Il1rl1*, *Tnfrsf4*, *Rora* were not differentially expressed between skin and colon Treg NLT cells (Figure 2.2H, Figure A.2A). Despite their resemblance, differences in function and/or state between skin and colon Treg NLT might reside in a few genes. Among these are *Dgat2*, an enzyme involved in lipid synthesis in skin (Fagerberg et al., 2014), and *Ikzf4*, a transcription factor relevant for Treg stability (Sharma et al., 2013).

The same approach applied to Treg cells from the spleen, mLN and bLN (Figure 2.2G) classified most central and effector Treg cells as Treg LT-like cells. Treg NLT-like cells, on the other hand, were more similar to Treg NLT and Treg suppressive cells. Both the mLN and the bLN had a higher proportion of Treg cells assigned as suppressive than spleen, which contained the highest fraction of Treg NLT cells. We confirmed the presence and proportions of Treg cell subpopulations in the Smart-seq2 datasets by matching these cells to the subpopulations found across LTs and NLTs in the 10x dataset (Figure A.2B).

Clustering of Tmem cells revealed multiple subpopulations (T helper-1 (Th1 cell), Th2 cells, Th17 cells, T follicular helper (Tfh) cells, lymphoid) (Figure A.2C and D) distributed differently across the tissues analysed (Figure A.2D). Th1, Th2 and Th17 cells in lymphoid tissues exhibited a stronger NLT phenotype than Tmem lymphoid cells and Tfh cells (Figure A.2E), which is likely an indication of their ability to adapt to and function in the NLTs.

In summary, scRNA-seq allowed us to dissect the heterogeneity of Treg cells from LTs and NLTs. We identified NLT- and LT-like Treg cell subpopulations that suggest progressive cross-tissue adaptation to the NLT environment. We found a close correspondence between skin and colonic Treg NLT cells, whilst revealing differences in gene expression that might explain their adaptation to the two environments.

2.2.3 Treg cells adapting to skin and colon share a transcriptional trajectory

The mechanisms underlying Treg cell recruitment and adaptation from LT to NLT are far from understood. Having identified multiple subpopulations at different stages of NLT adaptation (Figure 2.2D), we further dissected the dynamics of this transition. We obtained evidence of $CD4^+$ T cell recruitment from LT to NLT by reconstructing TCR clonotypes using TraCeR (Stubbington et al., 2016) from the Smart-seq2 datasets. This showed Tmem and Treg cell clones present in LNs and respective NLTs (Figure A.4A and A.4B), suggesting cell migration between them.

To identify Treg cell LN-to-NLT adaptation trends in the data, we reconstructed a pseudospace relationship between cells by obtaining latent variables (LV) from Bayesian Gaussian Process Latent Variable Modelling (BGPLVM, see Methods) (Titsias and Lawrence, 2010). Along the mLN to colon trajectory laid out by LV0, Treg cells are ordered from cTreg to eTreg cells, followed by NLT-like and LT-like Treg cells, and ending with the overlapping Treg suppressive and Treg NLT cell subpopulations (Figure 2.3A, "Colon" density plot, Figure A.3A). This order matches the increasing expression of NLT marker genes and decrease of LT ones across mLN subpopulations (Figure 2.2B and D). Importantly, Treg NLT-like cells from the mLN partially mixed with Treg LT-like cells from the colon, supporting the notion that NLT adaptation is a continuous process spanning LT and NLT. Overall, LV0 accurately represented the progressive migration and adaptation of Treg cells to the NLT environment, providing a reference to study the gene expression dynamics along this process. Skin and bLN Treg cells were projected onto the latent space defined for colon and mLN, resulting in a similar subpopulation distribution (Figure 2.3A, "Skin" density plot; see Methods). Nevertheless, a similar projection was observed when using just those cells (Figure A.3A and B). Applying the same approach to the Smart-seq2 datasets yielded similar distributions of the inferred cell subpopulations (Figure A.2B) along the LT-to-NLT adaptation trajectory, as well as considerable overlaps between LV correlated genes (Figure A.3C-E). The use of velocyto (Manno et al., 2018) to infer the directionality of adaptation suggests that most Treg cells found in the NLTs, as well as some of the NLT-like Treg and eTreg cells, are adapting towards a more pronounced NLT phenotype (Figure A.3C).

We then used the inferred LN-NLT trajectory to identify the cascade of transcriptional changes driving adaptation to NLTs by modelling genes with a sigmoid curve and find their activation or deactivation "times" (Figure 2.3B; see Methods). We

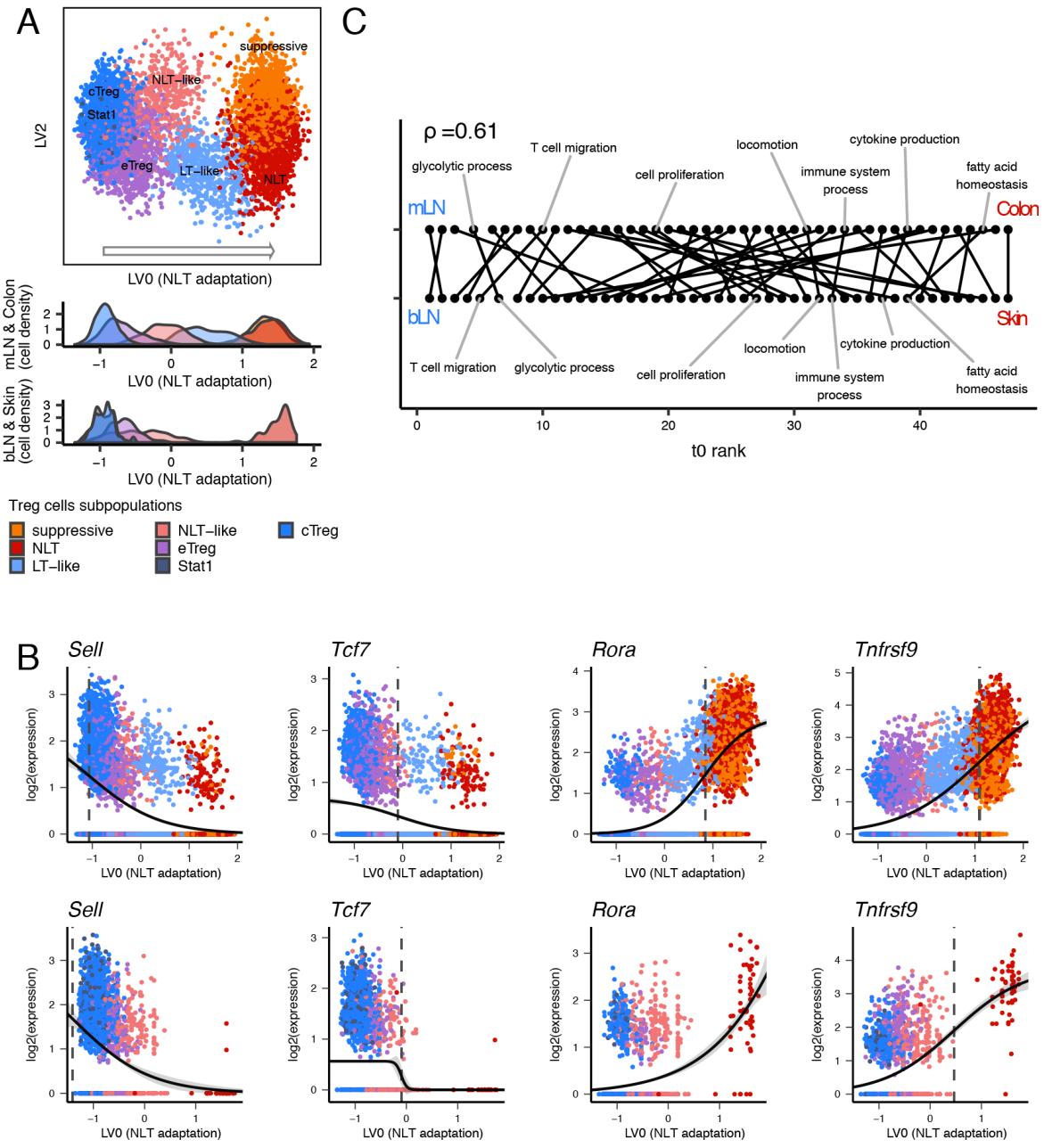


Fig. 2.3: **Reconstruction of Treg cell recruitment from lymphoid to non-lymphoid tissues in steady-state**
(A) Top two latent variables (LV) found with BGPLVM for mLN and colonic Treg cells, with bLN and skin Treg cells mapped over the same coordinates. **(B)** Gene expression in mLN and colon (top) or bLN and skin (bottom) over LV0 modelled as a sigmoidal curve. Dashed vertical line marks the activation point of each gene. **(C)** Sequence of activation of GO biological processes across the transition to colon (top) or skin (bottom), evidencing a conservation between both trajectories (Spearman's rho -0.61). See also Figures A.3 and A.4.

found 812 and 1209 genes with a switch in expression (either up or down) along the bLN-to-skin and mLN-to-colon trajectories, respectively, with 511 of those being shared. LT-related genes (*Lef1*, *Tcf7*, *Sell*) were downregulated, while NLT associated genes like *Nfil3*, *Ccr8*, *Cxcr6*, *Gzmb* were upregulated. TNFRSF-NF-κB-related genes (*Tnfrsf1b*, *Tnfrsf4*, *Tnfrsf18*) and the *Batf* transcription factor were upregulated still in the LN, reflecting the relevance of this pathway for eTreg cell development and the NLT phenotype (Vasanthakumar et al., 2017, 2015). Towards the NLT side of the trajectory there is evidence of further Treg cell differentiation, with upregulation of additional genes involved in this pathway (*Nfkb2*, *Tnfrsf9*), as well as other effector molecules (*Il10*, *Cd44*). Important regulators for the final tissue adaptation include *Rora*, recently described in skin Treg cells (Malhotra et al., 2018). We searched for enriched Biological Processes GO Terms, and calculated the mean time of activation or deactivation (t0) of the genes within each term. We found the gene expression kinetics along the adaptation trajectories to skin and to colon to be consistent (Spearman's rho=0.61, Figure 2.3C): T cell migration and glycolytic process are among the earlier events in both colon and skin, followed by cell proliferation; cytokine production and fatty acid homeostasis emerge towards the end of the adaptation trajectory.

In summary, we determined a continuous trajectory aligning Treg cell subpopulations from bLN, mLN, skin and colon according to the stage of recruitment and adaptation to the NLT environments. Furthermore, the consistent ordering of gene expression programmes shows that gene kinetics leading to NLT adaptation follows a similar regulatory sequence in both bLN-to-skin and mLN-to-colon trajectories.

2.2.4 Treg cell recruitment into skin and melanoma relies on common mechanisms

To validate our findings in steady-state cells, we used a mouse melanoma model to investigate if Treg cell migration and adaptation trajectory to peripheral tissues could be recapitulated. Previous studies analysing human TCR repertoires (Plitas et al., 2016; Sherwood et al., 2013) have shown that tumour-Treg cells are likely to be recruited *de novo* from LTs and not from the adjacent NLT, despite exhibiting a phenotype similar to that of NLT Treg cells (De Simone et al., 2016; Plitas et al., 2016). We therefore purified Treg and Tmem cells from B16.F10 melanomas or PBS controls 11 days after subcutaneous implantation into Foxp3-IRES-eGFP reporter

mice (Haribhai et al., 2007) to produce a plate-based scRNA-seq dataset (Figure 2.4A; see Methods).

Skin and tumour Treg cells clustered separately (Figure 2.4B). As with steady-state skin, we observed shared clonotypes between tumour and bLN Treg cells (Figure A.5B). In the tumour-bearing mice, we detected an additional cluster of cycling cells in both the LN and tumour (Figure A.5A). These observations suggest *de novo* recruitment from LN and simultaneous expansion in both tumour and draining-LN. DE between non-cycling tumour Treg and control skin Treg cells revealed a relatively small number of genes significantly different between the two Treg cell populations (28 upregulated in tumour and 10 in steady-state skin (Figure 2.4C)), in line with recently published human data (Plitas et al., 2016). Tumour Treg cells upregulate the exhaustion marker *Lag3* (Malik et al., 2017), as well as *Cxcr3* and *Ccl5*, while control skin Treg cells upregulate skin Treg cell markers such as *Il1rl1*, *Pim1*, *Sdc4*, *Kdm6b* and *Erdr1*. However, skin Treg cell signature genes such as *Batf*, *Tnfrsf4*, *Tnfrsf9*, *Samsn1*, *Tigit*, *Tchp*, *Ccr8*, *Ccr2* and *Itgav* are similarly expressed in both populations.

Next, we sought to obtain a shared migration trajectory of steady-state versus perturbed system (tumour model) Treg NLT cells recruitment. To this end, we used the MRD-BGPLVM algorithm (Damianou et al., 2012) (see Methods) to explore gene expression trends across Treg cells from the control skin, tumour and respective draining-LNs together. Two main latent variables were identified, one explained almost entirely by cell-cycle-associated variability (LV5), and one mainly associated with the LT-NLT signature (LV9) (Figure 2.4D, Figure A.5C). Notably, NLT adaptation trajectory (LV9) was strongly related to the trajectories found in control and melanoma conditions when MRD-BGPLVM is applied to each one individually (respectively, 86% and 61% of genes correlated with LV9 are also correlated with control LV1 and tumour LV1; Figure A.5E-H, see Methods).

Gene kinetics along NLT adaptation (LV9) for each condition show 158 shared genes, with 71% of which also present in the steady-state skin trajectory determined previously. Values of t0 remain largely unchanged between control and melanoma (Figure 2.4E), suggesting that NLT recruitment and adaptation follow the same program in homeostatic and perturbed conditions. The tissue adaptation genes shared between control and melanoma include many of the players in the TNFRSF-NF-κB pathway we previously described in the steady-state (*Tnfrsf9*, *Tnfrsf18*). These were accompanied by genes associated with cell migration and adhesion (*Ccr2*, *Gpr55*, *Plxna2*), transcription factors (*Rora*, *Ikzf3*, *Id2*, *Batf*, *Hif1a*, *Prdm1*), secreted factors

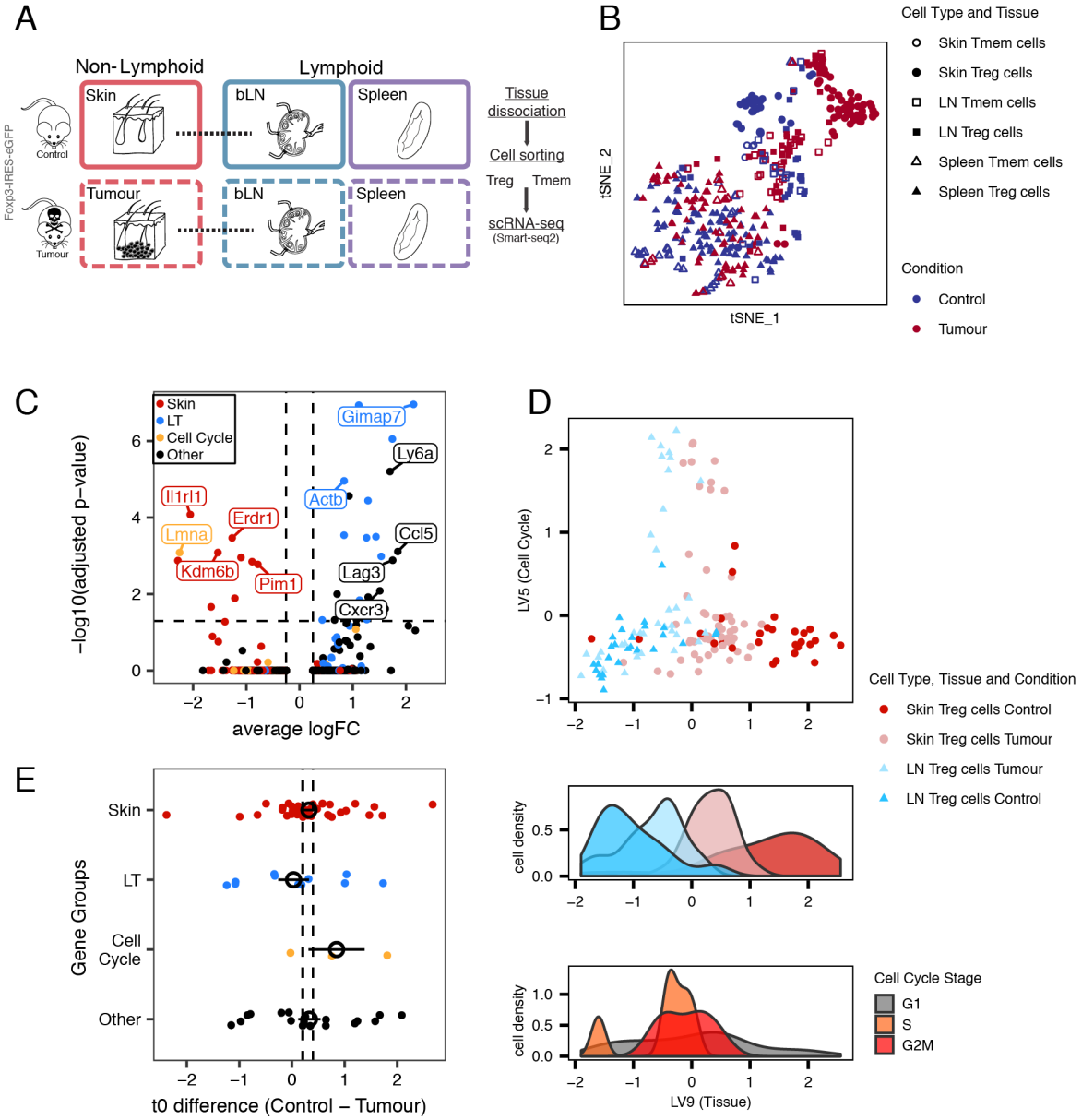


Fig. 2.4: **Recruitment and adaptation of Treg cells to the tumour environment recapitulates steady-state migration**
(A) Melanoma induction strategy and sampled tissues. **(B)** t-SNE depicting Treg and Tmem cells from tumour and steady-state skin, draining brachial lymph nodes and spleen. **(C)** Differential expression between skin and tumour Treg cells. Treg cells classified as cycling were excluded. **(D)** (top) Latent variables found with MRD-BGPLVM representing cell cycle (LV5) and non-lymphoid tissue recruitment/adaptation of Treg cells (LV9). (bottom) Distribution of cells based on Tissue and Condition and Cell Cycle phase along the recruitment trajectory. (Continued on the following page.)

Fig. 2.4: (continued) **(E)** Difference in activation time (t0) of genes in control and tumour. Genes are classified as being markers of skin, lymph node, cell cycle or other. Coloured points show mean +/- mean standard error for each group. Vertical dashed lines represent the mean +/- standard error for all t0 values. T-test between control and melanoma t0 indicates no change (p-value = 0.2631), with t0 values having a Spearman correlation coefficient of 0.65 between both conditions. See also Figure A.5.

(*Lgasl1*), and others related to immune activation and effector states (*Klrg1*, *Icos*, *Tigit*, *Gzmb*).

Despite the similarities between melanoma and control trajectories, cells from both conditions do not completely overlap, and Treg cells could be ordered by NLT adaptation between populations (from least to most adapted cells: control LN, melanoma LN, tumour, and control skin) (Figure 2.4D). This implies that in response to an immune challenge in a barrier tissue, a higher fraction of Treg cells in the LNs acquires NLT adaptations. In fact, for several NLT markers we observed more cells expressing them in the tumour-draining LN compared to the control, e.g. *Id2* (59% vs 26%), *Batf* (57% vs 26%), *Lgals1* (89% vs 67%), further supporting our hypothesis that there is priming of Treg cells to NLTs while still in the LN. Overall, Treg cells from challenged mice recapitulate the steady-state NLT adaptation.

2.2.5 Conserved NLT identity in mouse and human

We complemented our characterisation of murine NLT Treg and Tmem cells by collecting human Treg cells, as well as Tmem (sorted into central and effector memory) cells from blood and skin, and from tumour-adjacent colon sections from patients undergoing colonic resection (Figure 2.5A, Figure A.6). Similar to the mouse analysis, we identified gene markers for human $CD4^+$ T cell populations (see Methods).

Focusing on one-to-one orthologs, we found that 24 out of 144 human skin Treg cell markers and 17 out of 74 human colon Treg cell markers overlapped with the respective mouse signature. In colon, we observe the conservation of *Tnfrsf4*, *Lgals1*, *Srgn*, *Cxcr6*, *Maf*, or *Ikzf3* (Figure 2.5B), genes that we had previously identified as important in defining tissue identity and Treg cell subpopulations. The same applied to skin Treg cells, where we saw expression of *Batf*, *Rora*, *Rel*, *Srgn*, *Tnfrsf18*, and *Tigit* across species (Figure 2.5C). Overall, this indicates a conserved role of the core NLT signature, namely the TNFRSF-NF-κB-pathway.

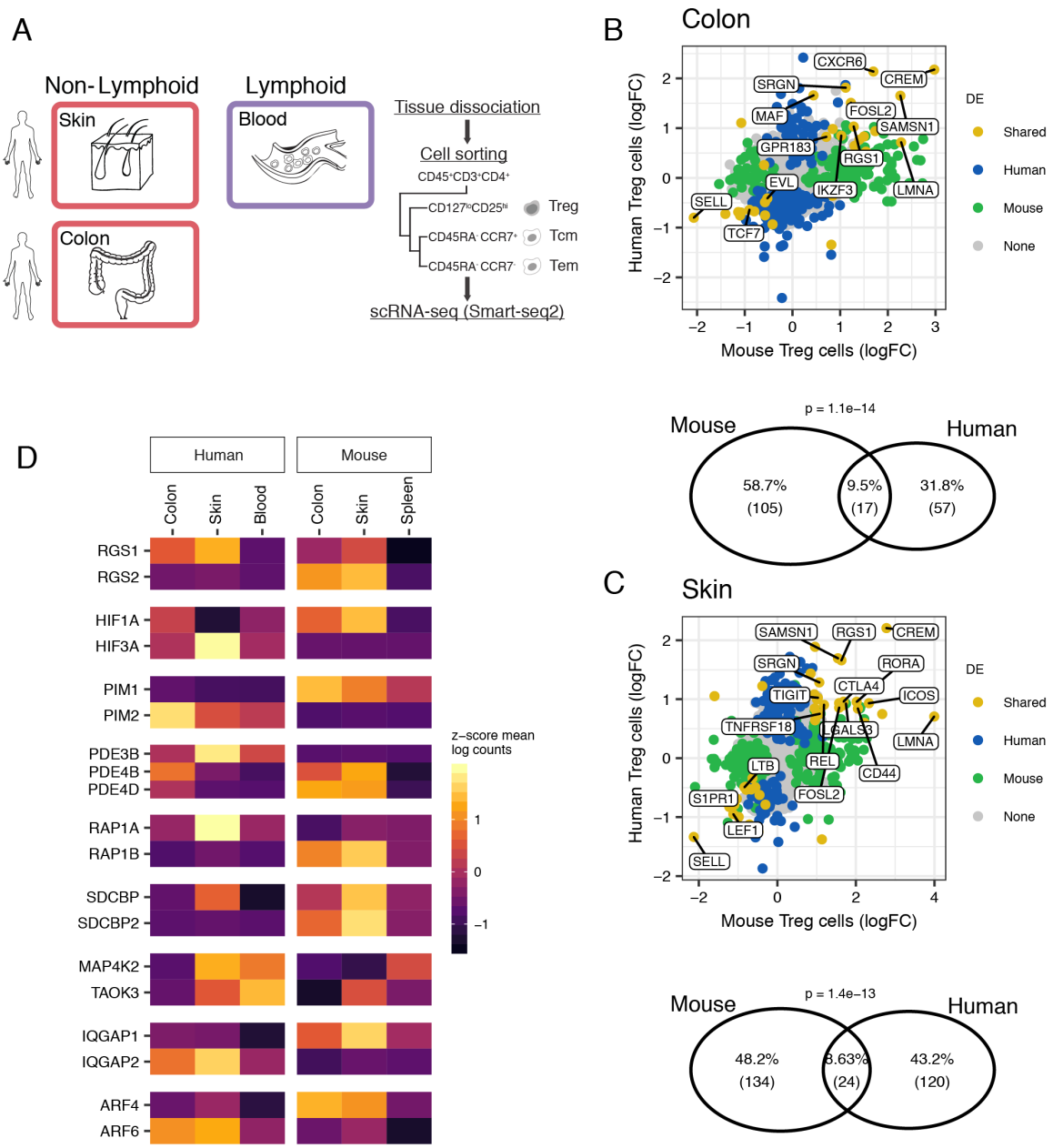


Fig. 2.5: **Human-mouse comparison of NLT Treg cell marker genes.**
(A) Tissues and cell types sampled from human. **(B and C)** Top: Overlap between NLT Treg cell markers detected in human and mouse, in either (B) colon or (C) skin datasets. Bottom: Fold-change between gene expression in non-lymphoid and lymphoid tissues in mouse and human. Blood and spleen were used as lymphoid tissues in human and mouse respectively. **(D)** NLT paralogs exhibiting opposing expression patterns between human and mouse. See also Figure A.6.

In several instances we observed the expression pattern of one gene being substituted by a paralog in the other organism (Figure 2.5D). For example, while the kinase *Pim1* is a marker of mouse NLT Treg cells, and was not expressed in human, the inverse was true of *Pim2*. A similar situation was observed for *Rgs1-Rgs2*, *Hif1a-Hif3a* and others. This suggests that some paralogous proteins have evolved to substitute each other during evolution of NLT Treg cells in mammals. The fact that several of the identified cases are receptors related to signal transduction leads us to believe that evolution of cell-cell communication pathways owes some plasticity to differential paralog usage.

Our cross-species comparison suggests that despite cross-species differences, the NLT Treg cell adaptation program defined in mouse is generally conserved in human.

2.2.6 Classfication of Treg cell populations across species

The increase in the number of datasets from a variety of species enables the comparison of cell types between them. In the previous section, it was explained how the core NLT Treg cell programmes were conserved between mouse and human, based on one-to-one orthologs. This type of orthologs make up the majority of conserved genes between these species, but one-to-many and many-to-many orthologs can also have important roles in cell identity and function.

Two logistic regression models were trained to detect mouse Treg cell subpopulations in human cells (see Methods). The first model was trained solely using one-to-one orthologs expressed in both datasets. The second model included all genes with any sort of orthology by adding the counts of related genes. Within each tissue, predicted subpopulations appeared as expected, with blood containing more cTreg and eTreg cells than NLTs, which in their turn had more Treg NLT and suppressive cells (Figure 2.6A). Transition populations (Treg LT-like and NLT-like) appeared more represented in general, as well as present in both lymphoid and non-lymphoid tissues. This is an effect comparable to that observed in Figure 2.2G, where Treg NLT-like cells are shown to match Treg NLT or Treg LT-like cells from colon, likely because of the intermediate phenotype of the cells.

When comparing the models per tissue (Figure 2.6A, top vs bottom row), similar subpopulation proportions are predicted per human tissue, indicating reduced differences between methods. However, we observe that the "one-to-one orthologs" model is the only that unexpectedly predicts the presence of Treg NLT in blood, and predicts in general a higher number of the rare, bLN-restricted Treg Stat1 subpop-

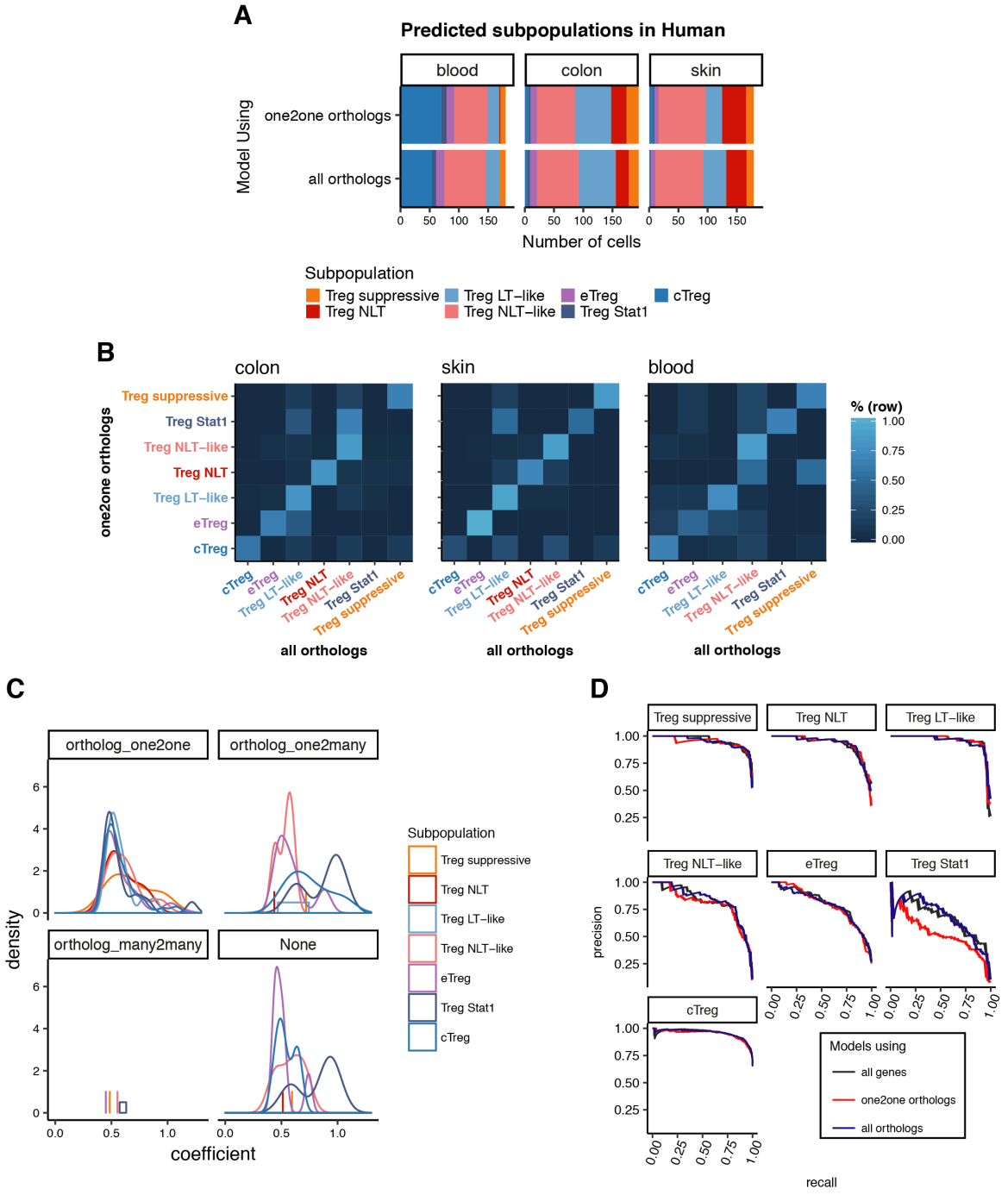


Fig. 2.6: **Training models for cross-species Treg classification.**
(A) Treg cells of each human tissue classified as each subpopulation detected in mouse using a logistic regression model trained with one-to-one orthologs (top) or all orthologs (bottom). **(B)** Row-normalised confusion matrices for each tissue, comparing the classifications using the one-to-one ortholog model (y-axis) against the all orthologs model (x-axis). (Continued on the following page.)

Fig. 2.6: (continued) **(C)** Distribution of the absolute value of coefficients for the top 300 genes learned for each population, in a model using all expressed mouse genes. **(D)** Precision-recall curves for models trained using all mouse genes, all orthologs or just one-to-one mouse-human orthologs. Precision and recall were calculated on a balanced test set composed of 10% of mouse Treg cells.

ulation (Figure 2.6B). In the "all orthologs" model, cells that would be assigned to this subpopulation are instead distributed between Treg NLT-like or LT-like, which are more evidently present in the same tissues in mouse.

To examine the contribution of different types of genes for cell identity prediction, we used a model trained on all mouse genes, and plotted the genes with the top 300 coefficients in absolute value by subpopulation and orthology type (Figure 2.6C. This shows that, while different subpopulations have similar distributions of absolute coefficients for one-to-one ortholog genes, distributions for one-to-many orthologs (and genes with no listed ortholog) are more dissimilar. In particular, Treg Stat1 cells have a larger number of one-to-many ortholog genes with a higher coefficient than the remaining populations, underscoring the importance of this type of genes in defining this subset. Concomitantly, precision-recall curves calculated for a test set comprised of 10% of mouse Treg shows that, while most subpopulations are equally well classified by both ortholog-based models, for Treg Stat1 cells only the "all orthologs" model performs as well as the "all genes" full model. These observations provide evidence that, while most cell states can be distinguished from one-to-one orthologs alone, this may not always be the case.

2.3 Discussion

Our work sheds light on the phenotype of skin and colon Treg cells. We profiled NLT Treg and Tmem cells to identify global relationships between cell populations, discriminating general $CD4^+$ and specific Treg cell markers in NLT. We found that these Treg populations conserve fundamental traits shared across the skin and colon compartments, namely a substantial prevalence of genes part of the TNFRSF-NF-κB axis. We leveraged the single-cell resolution of our data to explain Treg cell heterogeneity in the context of LT-to-NLT transition. Besides the eTreg cell state previously described in lymphoid organs (Cretney et al., 2011), we found two transitional subpopulations, Treg NLT-like cells in the lymphoid tissues and Treg LT-like cell in the non-lymphoid ones, which together explain the cross-tissue transition from central Treg to Treg NLT cell populations. NLT-like Treg cells in the mLN and bLN

showed extensive NLT-priming, including the upregulation of tissue-specific homing-molecules to the drained NLT. Others have demonstrated that a subpopulation of spleen Treg cells can express a partial visceral adipose tissue (VAT) signature and later give rise to fully-mature VAT-Treg cells upon migration (Li et al., 2018a), implying that this is valid for various tissues and should be considered in the design of future precision medicine strategies involving targeting of Treg cells to NLTs.

Comparative analysis of Treg cell phenotypes revealed genes associated with the TNFRSF-NF-κB axis to be highly upregulated in NLT (Figure 2.1C). Further enrichment analysis (Figure 2.7) confirms that this pathway is significantly associated with an NLT phenotype, despite the incomplete nature of the pathway's annotation. Genes such as *Tnfrsf4* and *Tnfrsf9*, which encode for receptors that play a role in inhibiting Treg cell function (Nagar et al., 2010), were identified here as distinctively associated with NLT Treg cells, yet are absent from the currently available pathway annotation. Mining of the dataset presented here can shed light on tissue-specific Treg cell biology, and reveal additional targets for Treg cell modulation.

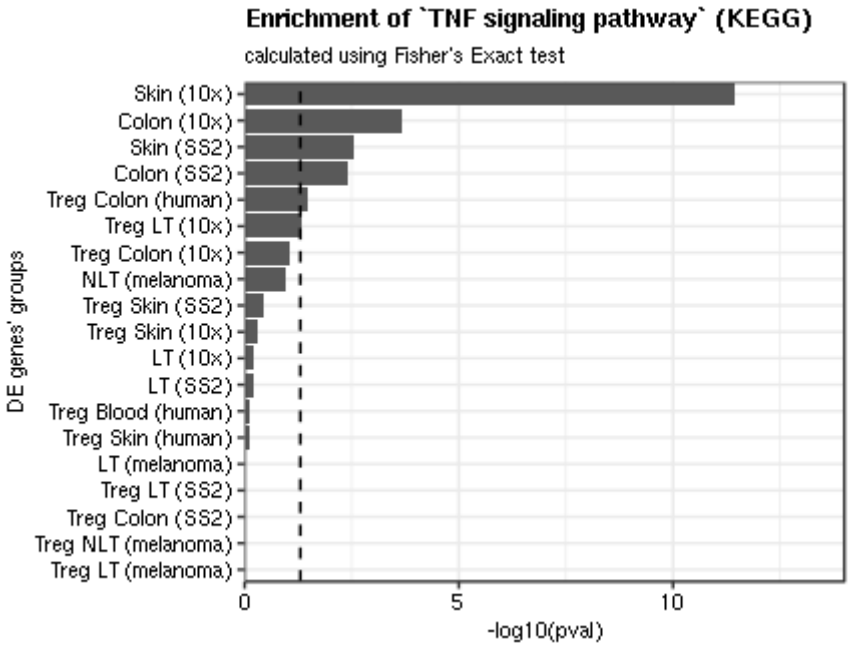


Fig. 2.7: **Enrichment of genes from the TNF pathway in NLT T cells.**
Barplot shows -log10(p-value) from Fisher's Exact Test, testing the overrepresentation of genes from the *TNF signaling pathway*, taken from the KEGG Database. Treg cell gene sets correspond to the intersection of genes upregulated in Treg cells intersected with those upregulated in the specific tissue (coloured dots in Figure 1.1C).

Our pseudotime results support migration and adaptation relationships between subpopulations, and allowed us to explore the basic mechanisms for the establishment of peripheral Treg cell phenotypes. In this transition, metabolic and proliferation changes in Treg cells happen concurrently with priming for migration, followed by changes in cytokine production machinery upon establishment in the periphery. Despite the overall similarity of recruitment and adaptation to NLTs, and although all three subpopulations (skin NLT, colon NLT, colon suppressive) fell close along the NLT adaptation trajectory, colon but mainly skin Treg NLT cells exhibited greater adaptation to the NLT environment. We hypothesise that the upregulation of *Ikzf4*, *Dgat2* and *Itgae* observed in skin might explain and contribute to the further stabilisation, retention and metabolic adaptation of Treg cells to the NLT compartment.

Treg cell priming in LNs is apparent from their increased NLT signature and expression of tissue-homing molecules, yet it is likely that Treg NLT-like cells are a heterogeneous subpopulation, with some cells egressing to the NLTs and others recently drained from the NLTs. This was confirmed using velocyto, and agrees with the bidirectional migration between LNs and the NLTs described in skin using a photoconversion system (Matsushima and Takashima, 2010). Studies coupling photoconversion and scRNA-seq can further our understanding of Treg cell migration patterns, as previously shown with single-cell qPCR (Ikebuchi et al., 2016).

A considerable proportion of the adaptation programme between bLN-to-tumour was contained within the bLN-to-skin trajectories. Similarly to steady-state, cues derived from NLTs are likely to prime Treg cells located in the draining LNs, as indicated by a higher percentage of cells expressing *Batf*, *Lgals1*, *Id2* and other NLT markers in melanoma. In sum, tumour Treg cells resemble less mature versions of their homeostatic skin counterparts that, nevertheless, follow the same NLT adaptation trajectory.

The establishment of correct orthology relationships can be important for cross species comparisons. While we show that including a broader variety of ortholog genes improves prediction for one Treg subpopulation, this is not a definitive solution and should warrant further testing. A drawback still present is the exclusion of genes with no defined orthology relationship. These could be included by an approach that aggregated the genes by gene sets that would match between species, which can be agnostic to these evolutionary relationships and instead rely on per-species gene functional descriptions. It can however leave out less well studied genes, or have poorer performance for less well described or annotated species.

Despite the conserved tissue-specific signatures, the differential paralog usage identified between species (Figure 2.5D) suggests a pivotal role for expanded gene families in rewiring signalling pathways throughout evolution. Studies focusing not only on tissue-resident cells, but also on the surrounding environment and organs can help dissect the relevance of these pathways in T cell biology, and how this evolutionary rewiring might affect immune response and homeostasis.

Overall, we reveal a dynamic adaptation of T cells as they traffic across tissues, and provide an open resource (data.teichlab.org) for investigating *in vivo* $CD4^+$ T cell phenotypes in mouse and human, to ultimately harness NLT $CD4^+$ T cells as future therapeutic targets.

2.4 Methods

For further experimental methods see Appendix A.

2.4.1 RNA expression quantification and normalisation

Sequencing data from 10x runs was aligned and quantified using the CellRanger software package with default parameters.

Gene expression from Smart-seq2 scRNA-seq data was quantified in counts using Salmon v0.6.0 (Patro et al., 2017), with the parameters –fldMax 150000000 –fldMean 350 –fldSD 250 –numBootstraps 100 –biasCorrect –allowOrphans –useVBOpt. For mouse, the cDNA sequences used contain genes from GRCm38 and sequences from RepBase, as well as ERCC sequences and an EGFP sequence. Since the EGFP RNA is transcribed together with *Foxp3*, counts from these two genes were added after quantification to represent *Foxp3* expression. For human data quantification, cDNA sequences from GRCh38 and ERCC were used.

Standard scRNA-seq analysis (QC, differential expression and marker gene detection, and clustering) was performed using Seurat (Satija et al., 2015). All data was log-normalised using the NormaliseData function with a scale factor of 10000. Our expression data for different tissues is also available for user-friendly interactive browsing online at data.teichlab.org.

2.4.2 scRNA-seq quality control

Quality control of 10x-derived data was made taking into account number of UMIs - keeping cells with between 1000 and 15000 UMI - and number of genes - keeping cells with between 700 and 3500 genes with at least 1 UMI (Table A.5). While cells were not filtered by their mitochondrial read content, cells with an elevated number of these reads are eventually removed via clustering (see "Subpopulation detection in 10x data").

For Smart-seq2 data, count values for each cell were grouped in an expression matrix, and ERCC expression were separated from true gene expression. Cells were then filtered based on different quality parameters calculated for each dataset (Table A.5). Additionally, the output of TraCeR (Stubbington et al., 2016) was used to remove cells without a detected TCR sequence, as well as invariant Natural Killer T (iNKT) cells and $\gamma\delta$ T cells (defined as cells with at least one γ and one δ chain detected and no $\alpha\beta$ pair). For the colon and skin datasets, 433 and 745 cells passed quality control, respectively.

Importantly, we note that TCR detection greatly improved our filtering by excluding cell types captured by FACS that did not fit the intended categories. This is the case for iNKT cells - captured mostly together with spleen T memory cells - and $\gamma\delta$-T cells - sorted together with skin Tmem cells in the melanoma experiment. Indeed, we also identified a NKT population in the 10x dataset, mostly within the cells sorted as spleen Tmem cells, as well as some LN Tmem cells (Figure A.1B and A.1C). We cannot, however, state that these are "invariant", since we have no access to their complete TCR chains. TCR filtering also enables removal of cell doublets by identifying cells expressing an excessive diversity of recombined TCR chains. Even in cases of no allelic exclusion for TCR α and β sequences, each cell would still only be able to produce two recombinants of each, allowing removal of cell doublets expressing more than two recombinants for a TCR locus. Lastly, we removed all cells not expressing any recombinant TCR in order to have a more stringent quality control. While in the human dataset the number of cells without a TCR was evenly distributed across tissues and cell types, there was a clear skew towards TCR absence in peripheral Treg cells (colon and skin) in the mouse datasets. These Treg cells did not appear to differ from the remaining population, having no differentially expressed genes or major differences in their overall number, presenting only a skew towards a higher number of reads (data not shown).

2.4.3 Dimensionality reduction methods

To obtain an overview of the datasets showing the relationships between cell population clusters, Principal Component Analysis (PCA) and tSNE were used. Before PCA, data was scaled using the ScaleData function (negative binomial model, normalising by the number of UMI and centering the data). PCA and tSNE were calculated using the RunPCA and RunTSNE functions, respectively. For each dataset, a different number of Principal Components (PCs) and values for perplexity were used (Table A.5), chosen by visual inspection of an elbow plot representing the relative importance of each PC. With exception of the PCA projection for the complete 10x dataset, only highly variable genes were used, calculated using the FindVariableGenes function from Seurat with the parameter 'num.bin' of 100 and 'binning.method' of "equal_frequency". Using all genes for dimensionality reduction of the whole 10x dataset resulted in more accurate clustering, allowing for the identification of most contaminant cells on this first step (Figure A.1B). Plate-based datasets were treated separately as much as possible to avoid confounding batch effects from experiments performed separately.

2.4.4 Subpopulation detection in 10x data

To find clusters in the data, we used the FindClusters function from Seurat, with the same number of principal components used for tSNE. Cluster annotation was done by inspecting markers detected by the FindAllMarkers function.

Global clustering of the 10x dataset was done with the resolution parameter set to 0.2. After clustering the complete dataset, we excluded artifactual subpopulations (Figure A.1). A mixed Treg and Tmem cell population characterised by high expression of immediate-early response genes (e.g. Jun, Junb, Fos, Fosb), which has previously been reported in other cell types (Adam et al., 2017; van den Brink et al., 2017; Wu et al., 2017) was removed. An additional population of lymphoid tissue Tmem cells was also excluded because they presented expression profiles similar to NKT cells (*Nkg7*, *Ccl5*, *Cd160*, *Klrbc1*, *Cxcr6*).

Clustering on individual tissues used the following resolutions: for Treg cells, 0.3 on Spleen, 0.4 on bLN, 0.4 on mLN, 0.5 on Colon, 0.4 on all skin cells; for Tmem cells 0.4 on Spleen, 0.3 on bLN, 0.7 on mLN, and 0.6 on Colon. Annotation was performed and subpopulations characterised by immediate-early response genes were removed.

2.4.5 Differential expression analysis

Differential expression (DE) and marker gene detection was performed using the FindMarkers and the FindaAllMarkers functions from the Seurat R package, using the default Wilcoxon test. Genes were considered differentially expressed if they had an average log fold-change of at least 0.25 and a Bonferroni-adjusted p-value of 0.05 or lower.

For DE including all cells of the 10x dataset, a minimum of 5% of cells had to express the gene, otherwise a minimum of 1% was used. For comparisons between tests (for example Treg vs Tmem cells and LT vs NLT, see Figure 2.1C), the FindMarkers function was run twice - the first time to determine all genes considered expressed for each comparison, the second using the union of all those genes.

In the human and mouse comparison, human NLTs were compared to blood and mouse NLTs were compared to spleen only, and testing was restricted to genes with one-to-one orthologs.

2.4.6 Mapping cells to known populations using logistic regression classification

To make a correspondence of cells in the 10x dataset with the identified Treg cell subtypes in the colon (Figure 2.2G), or between Smart-seq2 data and the complete 10x dataset (Figure A.2B), the counts and subpopulation labels of the 10x dataset Treg cell subpopulations and the complete 10x dataset were used to train a logistic regression classification model using scikit-learn with an L1 penalty and default parameters. The label with the highest probability predicted by the model was then attributed to each cell. The figures show the percentage of each tested population that was predicted as matching to each learned label.

For cross-species mapping of Treg subpopulations, 90% of the sorted Treg cells from mouse were used to construct two models, with the remaining subpopulation-balanced partition kept separately for model testing. The first model (refered to in Figure 2.6 as "one-to-one orthologs") was used only using genes expressed in both species that are one-to-one orthologs. Another model was trained by using all genes with known orthologs, and adding the counts for genes with many orthologs. For example, if a gene in mouse corresponds to three genes in human (i.e. a one-to-many relationship), then the counts of the three human genes are added and given one identifier. For many-to-many relationships, the same happens in both species simultaneously. Additionally, a third model was trained using all mouse genes, to use

as a ground truth for the predictive power of the other models. With 10-fold cross validation, these three models have a mean accuracy of 84.0%, 84.7%, and 85.6%, respectively. Precision-recall curves were then calculated using the 10% test set.

2.4.7 Obtaining a migration latent variable for steady-state Treg cells

The large dimensionality of single-cell RNA-seq data has been used before to gain insights on time-dependent events (Lönnberg et al., 2017; Trapnell et al., 2014) by applying methods for pseudotime inference. Although it is impossible to follow one cell through the complete process, these methods can order single-cell data into a continuous dimension, using the discrete samples as snapshots containing a multitude of intermediate states.

Immune cells are expected to migrate between LTs and NLTs. We assumed that this effect would be reflected as a gradual single-cell expression phenotype, which could be captured as a latent variable of the data. To achieve this, we used Bayesian Gaussian Process Latent Variable Modelling (BGPLVM) (Titsias and Lawrence, 2010), implemented in the python package GPy (https://github.com/SheffieldML/GPy) as "GPy.models.BayesianGPLVM", which was already used before for dimensionality reduction in scRNA-seq data to model Th1-Tfh cell differentiation (Lönnberg et al., 2017). BGPLVM was used on log-scaled counts and only considering highly variable genes. We run the method with six latent variables (LV) to be sure we capture the most relevant ones by Automatic Relevance Determination (ARD, Figure A.3A), although this number does not alter significantly the performance of the algorithm. We then interpret the most important LV as the one ordering the cells between tissues along a migration and adaptation transition. In agreement, we observe gene expression changes associated with losing the lymphoid tissue identity and acquiring a peripheral tissue transcriptional profile (Figure 2.3B).

For 10x data, the method was used on mLN and colon Treg cells. We then mapped bLN and skin Treg cells onto the same LV using the predict function from the BGPLVM module, in order to have a similar coordinate system for both trajectories. Running BGPLVM with all data together would achieve a similar result (not shown). A BGPLVM projection of bLN and skin Treg cells (Figure A.3B) shows an identical projection but with a wider gap between bLN and skin cells due to the large differences in cell numbers. We excluded spleen cells from this analysis to focus specifically on LN to NLT adaptation.

Similar effects are also observed in the corresponding Smart-seq2 cells (Figure A.4D). We then show that all the LVs chosen as a "pseudospace variable" (LV0) have a similar effect between datasets by comparing the shared proportions of genes correlated with each of them (Figure A.4E).

2.4.8 Identifying a common tissue migration trajectory in control and melanoma

Similarly to the steady-state, migration from the LN to the skin with a melanoma challenge is also expected. A common between-tissue Treg cell migration trajectory in control and melanoma conditions was obtained using Manifold Relevance Determination (Damianou et al., 2012) (MRD). MRD works by having an underlying BGPLVM model whose dimensions can be shared or private between sections of the data. Having the prior knowledge that a cell-cycle effect is present in the data (Figure A.5A) and with the goal of obtaining a LV explaining tissue recruitment in both conditions, the melanoma dataset was divided into three sections for input: one with the expression in all cell-cycle associated genes, one with marker genes for any tissue, and one with the remaining genes. The importance of each section in each latent variable is shown in the ARD plot (Figure A.5C). The model was run allowing for 12 LVs as output, and the one highly influenced by tissue-specific genes but not cell-cycle or other genes was used as a migration trajectory for both conditions (Figure 2.4D). The effects captured by these LVs can be observed in BGPLVM projections for the individual conditions (Figure A.5E-G).

2.4.9 Switch-like genes in the migration latent variable

Gene expression changes in a continuous trajectory can be interpreted as a series of switch-like events. These can be modeled using a sigmoid curve, described by the following equation:

$$S = \frac{2 \times \mu_0}{1 + \mathrm{e}^{-k(t-t_0)}} \tag{2.1}$$

where μ_0 is the mean expression between the sigmoid "on" and "off" states, t_0 is the point in which the switch in expression happens, and k defines the sigmoid inclination and can be interpreted as the activation strength. Parameter k will

additionally inform on the direction of the switch (activation or inhibition) from its signal.

The R package switchde (Campbell and Yau, 2017) was used to model gene expression as a sigmoid in the inferred migration trajectories, using the appropriate latent variable as pseudotime.

In the steady-state 10x dataset partitions (mLN+colon Treg cells and bLN+skin Treg cells), switchde was applied for non-Tmem cell specific genes expressed in at least 30 cells, as well as genes with an absolute correlation greater than 0.25 with the LV chosen for both partitions. Due to the large differences in the number of cells in the skin partition, we ran switchde 100 times on different subsamples of each Treg cell subpopulation matching the smallest subpopulation size (405 for the colon partition, 55 for the skin partition), and used the median values of the parameters for further analysis. For the melanoma dataset, genes expressed in at least 5 cells in both conditions were tested. Only genes with a q-value<=0.05 and that had a t_0 within the LV range were kept for further interpretation.

2.4.10 RNA velocity estimation

RNA velocity is a measure that leverages detection of spliced and unspliced transcripts to predict single-cell development directionality (Manno et al., 2018). We used velocyto to determine in which direction cells were changing in the cross-tissue adaptation trajectories. We have followed the python implementation of velocyto, and the code can be found in https://github.com/tomasgomes/Treg_analysis/blob/master/Velocyto.ipynb, where each of the runs is present.

2.4.11 Detection of expanded clonotypes

T cell receptor (TCR) sequences were reconstructed from single-cell RNA-seq data and used to infer clonality using TraCeR (Stubbington et al., 2016). We used TraCeR with the parameters –loci A B D G, –max_junc_len 120 to allow reconstruction of TCRα, TCRβ, TCRδ and TCRγ chains in each cell and to permit TCRγ chains with long CDR3 regions.

2.4.12 GO Term enrichment

To test for enriched GO Biological Processes or KEGG Pathways in gene sets, the gprofiler R package (Reimand et al., 2016) was used, with the option of moderate

hierarchical filtering enabled. No custom background was used (i.e. all genes with a GO Term annotation were considered). To determine the succession of Biological Processes GO Terms (Figure 2.3C), we used the approach above on all genes called DE by switchde, and plotted only the terms with at least two genes.

2.4.13 Cell-cycle analysis

To assess potential effects of cell-cycle in the interpretation of the scRNA-seq datasets, Cyclone (Scialdone et al., 2015) (implemented in the scran R package) was used on all datasets. Results for the mouse melanoma dataset (where a relevant cycling population exists) were projected on the tSNE (Figure A.5A). As the vast majority of cells was assigned to the default cell-cycle stage (G0/G1 in mouse, S in human), no cell-cycle correction was performed.

2.5 Conclusions and future work

This Chapter has elucidated the molecular makeup of Treg cells in their tissue context, and revealed the transcriptional transitions these cells undergo during adaptation to a new tissue environment. Deep characterisation of colon, skin, their draining lymph nodes and spleen revealed evident transcriptional heterogeneity, reflected in distinct subpopulations likely associated with different activation and cross-tissue transition stages. The full steady-state profile of Treg cells requires further sampling of more tissues. Skin Treg cells, because they are harder to extract, were not as deeply sampled, yet some heterogeneity could still be inferred (Figure 2.2G).

Increasing the number of profiled tissues holds the promise of revealing further tissue-specific subpopulations, allowing for a full map of Treg cell phenotypic regulation to be compiled. Recent work has showed that Treg heterogeneity is associated with TCR activation in colon and spleen (Zemmour et al., 2018), and further integration of gene expression and open chromatin data has shed light into Treg cell tissue-specific regulatory networks (DiSpirito et al., 2018). In particular, this last study places NF-κB-related transcription factors (*Nfkb1*, *Nfkb2*, *Rel*, *Relb*) within the colon-specific regulatory network, demonstrating the increased power in combining data from different tissues. Within the colon, the authors of the study also identify a subpopulation of cells they presume to be circulating (expressing *Ccr7*, similarly to LT-like Treg in Figure 2.2A), and they are capable to in addition distinguish between thymic and periphery-derived Treg cells. Various factors (tissue processing protocol,

single-cell isolation method, among others) can influence the detection of the genes driving these populations to explain why they were not detected in the data here presented. Nonetheless, comparing these studies shows how useful it can be combining scRNA-seq data obtained from different sources. Lastly, all the studies here described have mostly focused on mouse. Recent analysis in our lab (James et al., 2019) has showed that, from total immune cells extracted from human colon and mesenteric lymph nodes, the Treg cell subpopulations described in this Chapter can also be detected, further confirming the robustness of this finding.

When sampling diverse tissues, their physical processing is crucial, not just to obtain a comprehensive representation of the cells present, but also in the way that such extraction protocols can affect cellular phenotypes. It has been described that, due to some processing methods, cells can undergo transcriptional changes, with the activation of immediate early genes as a response to stress, or activation of genes encoding for heat-shock proteins (van den Brink et al., 2017). Importantly, some of the immediate early genes are also implicated in immune response, such as *Fos* and *Jun*. Furthermore, this effect can be cell type-dependent, additionally confounding the interpretation of said data. Mitigation of these effects has been achieved in the past by inhibiting transcription during tissue processing (Wu et al., 2017). While in the present work we avoided drawing excessive interpretations regarding genes involved in these pathways, future cross-tissue works should account for these effects. This should ideally happen at the biological material processing stage, since some of these genes can have *bona fide* functions within the tissue-specific context (Wheaton and Ciofani, 2019).

The inferred transcriptional trajectory (Figure 2.3A) offers a base model for what tissue adaptation of Treg cells during trafficking might resemble. This trajectory was inferred under the assumption that all cross-tissue intermediate states are represented, however this might not be the case. The overlap between the LT-like and NLT-like Treg subpopulations is encouraging, pointing at these being the intermediate state of this transition. Indeed, NLT-like Treg cells in lymph nodes expressed surface receptors known to direct cells to their specific tissues (Figure 2.2E). However, it can also be argued that a true transition stage would have to be captured "in transit", i.e. obtained from blood. This might be hard to achieve given the very low representation of these cells compared to other circulating lymphocytes. In a model organism, it could potentially be addressed by genetically modifying Treg cells to express a detectable marker upon exit of lymph nodes or NLT, if such regulatory mechanism is completely understood. A further aspect to consider is the directionality of cell

trafficking. Velocyto analysis (Figure A.3C) hinted at a LN-to-NLT transition, however it was not conclusive, also showing some NLT-like cells to be adapting into a lymph node identity. It is indeed possible that trafficking occurs in both directions, yet the association of this movement with the detected subpopulations will only be revealed by combining single-cell sequencing with lineage tracing, for example using adoptive transfer into specific tissues.

It is also expected that the data here produced and dissected serves as a platform for future functional studies on Treg identity. This has already been the case in (Wheaton and Ciofani, 2019), where the authors, starting from the tissue-specific expression of *Junb* captured in this dataset, validate the importance of this gene for adaptation of the Treg cell effector programme in the colon, through the use of Treg cell-specific knock-out of the gene. Gene knock-out studies can be very powerful to test the importance of the genes here revealed to impact Treg identity and adaptation. This can be combined with lineage tracing of these cells to quantify how affected cell trafficking is, or with functional assays to evaluate whether the gene is important for Treg suppressive function, for example. In humans, functional validation is more restricted due to ethic concerns, yet the gene lists here produced can also be cross-referenced with genes involved in autoimmune or tissue-specific pathologies, shedding light into the role of tissue-specific Treg cells in these diseases.

The unravelling of Treg cell heterogeneity also feeds into the more general topic of how cell types can be classified. Based on their transcriptional phenotype, it is apparent that the Treg cell subpopulations represent transient states. Despite being evidently different when examining each individual tissue (Figure 2.2A), they can actually appear similar when compared with other tissues and cell types (Figure 2.1). For this reason, the establishment of a cell type reference should firstly consider cells in their tissue context, and only then establish the similarities across these. It can however be debated how accurately these different states could be distinguished in the context of a broader cell type classification. Future methods might aim at representing transient cell states separately from the defined, central cell identity.

Chapter 3

Developing a method to integrate and classify cell types across tissues

The widespread adoption of single-cell sequencing technologies has revolutionised the molecular profiling of cells. The use of these methods allows us to understand the building blocks of tissues at an unprecedented resolution. As further human tissues are examined, a catalog of cell types and their gene expression profile can be compiled from published data.

This chapter outlines the development of a computational pipeline for cross-tissue integration of scRNA-seq data, illustrating the performance of individual steps in a controlled dataset. The pipeline is tuned to capture cell type similarities from annotated and non-annotated data between tissues, resulting in an unbiased gene expression reference for cell type identity. This reference can then be used to train *CellTypist*, a set of logistic regression classifiers capable of assigning cell type identity to newly produced scRNA-seq data. *CellTypist* is trained on the *Tabula Muris* dataset as a mouse reference, and on a large collection of tissue-derived human data.

This project was initially conceived together with Valentine Svensson while he was part of the Teichmann group. The human data collection and integration was performed with the assistance of Ni Huang. The mouse and human cell type references trained here are further analysed in Chapter 4.

3.1 Introduction

The growth of the scRNA-seq field is in part due to an increasing number of complex and detailed cellular census of individual tissues, often directly associated with large

consortia that aggregate these datasets and establish guidelines and collaborations to identify all cell types across an organism (Regev et al., 2017). Individually, these studies have provided crucial insights into cell biology. Nonetheless, the data generated can often be reused for new purposes, either on its own to extract new conclusions, or through combination or comparison with novel data.

To combine scRNA-seq datasets, batch correction and batch alignment methods seek to either correct gene expression values accounting for technical metadata (Buettner et al., 2017; Büttner et al., 2019; Haghverdi et al., 2018; Johnson et al., 2007; Ritchie et al., 2015), or place cells from different batches, technologies and datasets in a common manifold, allowing joint clustering and pseudotime analysis (Butler et al., 2018; Hie et al., 2019a; Korsunsky et al., 2018; Polański et al., 2019; Stuart et al., 2019). Conversely, comparison between scRNA-seq datasets aims to impart the knowledge gained from one dataset into another, usually through classification models. Various methods have been developed to compare cell types (and indeed other labels) across datasets (reviewed in Chapter 1.3, Table 1.2). A benchmark of 22 classification methods for scRNA-seq (Abdelaal et al., 2019) has revealed SVM-based classifiers as the top performing ones, capable of accurate cross-dataset classification and handling of all genes by using L2 regularisation. This study also echoed the findings of another recent study (Köhler et al., 2019) that demonstrated that deep learning methods do not outperform classical machine learning approaches, including SVM and logistic regression, for cell type classification.

Despite the high accuracy of there tools, which are highly effective at annotating new data using specific datasets, their scope will is limited to the dataset chosen as a reference and don't directly handle large collections of data. To address the need for a reference that can allow fast and automatic annotation across tissues, we have developed a pipeline for integration of scRNA-seq data obtained from a variety of tissues, which can then be used to build *CellTypist*, a global cell type classification method based on logistic regression classifiers. While previous work has been developed for well annotated mouse data using a neural network-based classifier (Alavi et al., 2018), *CellTypist* can leverage data with different annotations, focusing on providing broad identifiers for cells based on a reference from pooled data.

This chapter discusses the structure of the integration and classification pipeline, exploring its strengths and caveats, with each step performed on the *Tabula Muris* data (Various, 2018). Despite coming from a sole publication, this dataset has the advantage that it was generated in highly controlled conditions, spans 20 tissues,

and includes a detailed and robust cell type annotation to be used as a ground truth for the performance of each stage. The methodology outline is then used to train classifiers based on the *Tabula Muris*, as well as a collection of human data of close to 1.5 million cells. The training accuracy and bias of the models is assessed and discussed, with suggestions for further improvements.

CellTypist is further explored in Chapter 4, where its application will be tested for automatic classification. It will also be explored in terms of biological insights, in order to identify cross-tissue relationships and examine which genes the model deemed important to determine cell identity.

3.2 Methodology

3.2.1 Per-tissue clustering to approximate cell type annotations

Processing the data that will be used as a reference in *CellTypist* follows three major steps. First, the data collected follows a procedure for uniform per-tissue processing (Figure 3.1A). Next, the clusters determined in each tissue are matched across the whole dataset (Figure 3.2A). Lastly, the combined clusters are used as labels to train a logistic regression classifier using the complete data collection (Figure 3.4A).

Most scRNA-seq studies profile cellular heterogeneity in one specific biological sample, which often results in reporting the cell types or cell states present. While this is clearly displayed in the figures from these studies, this information in not always supplied in a machine-readable format, associated with either the raw sequencing reads or the quantified gene expression. Furthermore, most annotations do not follow a uniform, controlled vocabulary (Bard et al., 2005), and is often done at varying resolutions depending on the focus of the study or the breadth of the dataset.

The pipeline starts by splitting the collected data by tissues, grouping together data from different studies that profile the same body part, even if using distinct scRNA-seq protocols (Figure 3.1A). At this stage, data from each tissue is processed following a uniform workflow using scanpy (Wolf et al., 2018). Gene expression is normalised by their total counts and log transformed, and different datasets are batch aligned using BBKNN (Polański et al., 2019). Lastly, clustering at varying resolutions using the Leiden algorithm (Traag et al., 2019) is performed. This processing is executed to ensure that all tissues are similarly treated, regardless of the level of annotation, and to allow unannotated data to be included and bolster *CellTypist*'s training data.

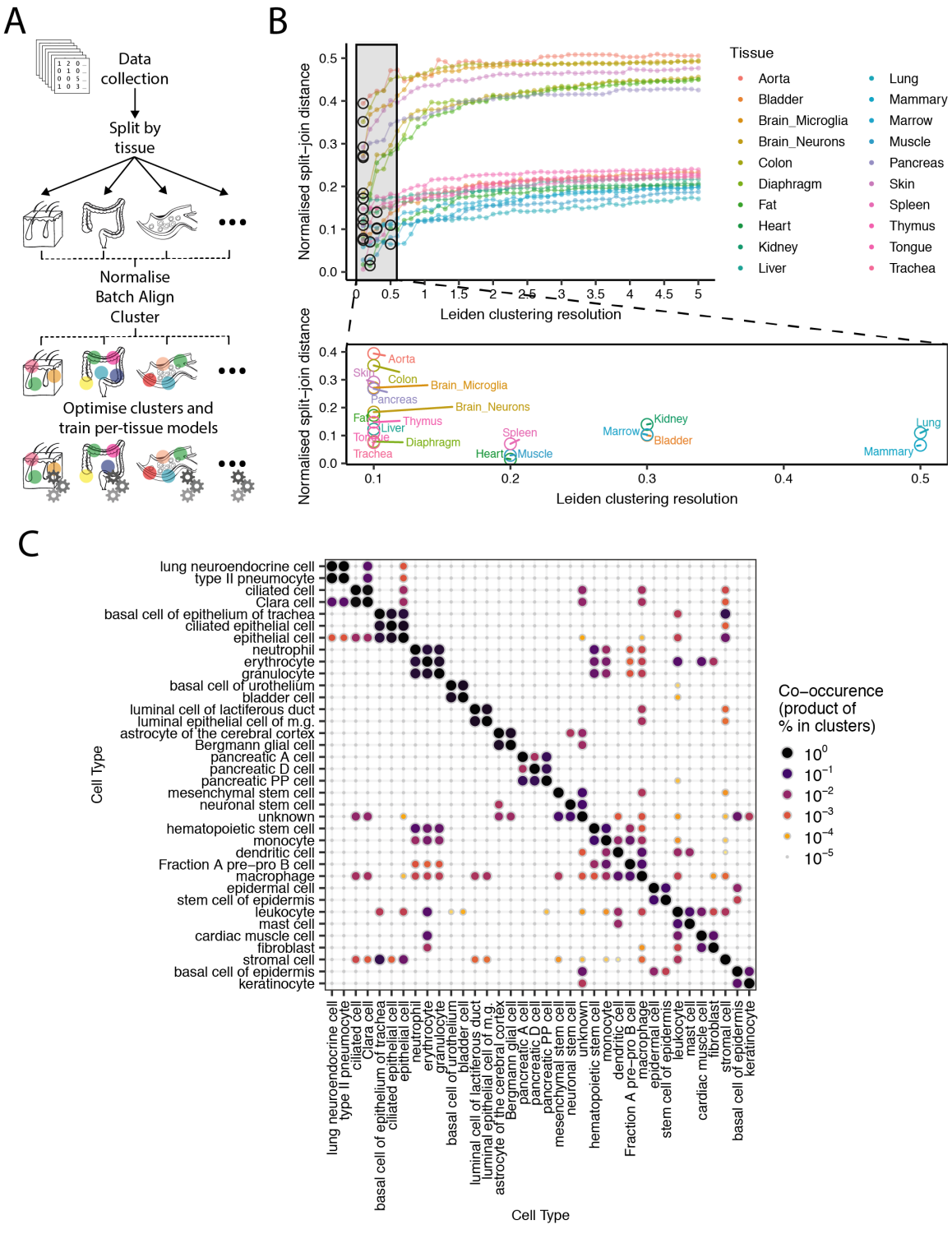


Fig. 3.1: **Data reprocessing per-tissue**
(A) Pipeline for initial data processing. Data collected is split into tissue, followed by integration of different datasets (in *Tabula Muris*, different protocols), and clustered to optimally match existing cell type annotations. (Continued on the following page.)

Fig. 3.1: (continued) **(B)** Per-tissue cluster optimisation, choosing the resolution that approximates existing cell type annotations. Similarity is measured with normalised split-join distance, and constrained to solutions with a number of clusters of at least as many as existing annotations in the tissue. Upper panel shows the full range of resolutions tested per-tissue; lower panel shows the resolution range in which the optimal value for each tissue was present. **(C)** Co-occurrence of annotated cell types in the same clusters, determined by summing the products of each cell types per-cluster percentage. Only cell types with at least one co-occurrence value of 0.05 were kept. m.g. - mammary gland.

The clustering performed in each tissue can be optimised to approximate existing cell type labels. To this end, various groupings were generated in each tissue using a range of values for the resolution parameter of the Leiden algorithm. The clusters obtained were subsequently compared to known cell type annotations. The distance between clusters and cell types was calculated using the normalised split-join (SJ) distance (Dongen, 2000). Briefly, SJ distance measures the distance between two data partitions according to the number of element-wise division (split) or merge (join) operations necessary to fully convert the new partition into the former. In this specific example, it counts the number of operations necessary to convert the new clustering groups back into any known cell type annotations. Since the values in the original metric are dependent on the number of elements being cluster, and which here differ between tissues, Figure 3.1B shows a normalised version of the metric, where it was divided by 2N, with N = number of cells for a given tissue. Its values will fall in an interval between 0 and 1, corresponding to complete similarity or complete dissimilarity.

The calculated normalised SJ distance for each Leiden clustering resolution tested is plotted in Figure 3.1B (top). The chosen resolutions (black circles) are a result of Leiden clustering that 1) outputs at least as many clusters as there are unique cell type labels in the largest dataset contributing to that tissue, and 2) has the lowest normalised split-join distance. Despite the broad range of clustering resolutions tested, (from 0.1 to 5 at 0.1 intervals), the parameters chosen for all tissues concentrated at resolutions of up to 0.5 (Figure 3.1B, top shaded box and bottom expansion). Tissues appeared to organise into two distinct groups: a smaller group with an SJ distance above 0.2, and a larger one with a distance below 0.2. Interestingly, all tissues in the higher SJ distance group were only sequenced using Smart-seq2 (Figure B.1), suggesting that integration of data from different protocols is not resulting in clusters that incorrectly approximate known cell types. This group also had, in general, fewer cell types annotated, thus the higher values could be explained by overclustering.

To understand the extent of misclustering of cell types within the per-tissue clustering step, co-clustering of known cell types was examined across all tissues (Figure 3.1C). We started with a cell type-by-cluster matrix, showing the distribution of cell types per cluster as a percentage. To obtain a symmetric matrix, i.e. show the co-occurrence of two cell types normalised for the occurrence of each cell type, we obtained the product of this matrix with its transposed form. The resulting matrix is subsequently filtered to only include cell types with at least one co-occurrence value of 0.05 (meaning more than 20% of the cells from each cell type in a pair would appear in the same clusters). Besides a high clustering of cells with the same annotation, the plot shows that most of the mixing happens between related cell types (e.g. epithelial cell, ciliated epithelial cell and basal cell of epithelium of trachea; luminal cell of lactiferous duct and luminal epithelial cell of mammary gland), or between hematopoietic-derived cells (e.g. neutrophil, erythrocyte, granulocyte; hematopoietic stem cell, macrophage and dendritic cell; leukocyte and mast cell). This is expected due to the similarities between related cell types, as well as due to less resolved clustering of immune and non-immune cells when these are analysed together. Another possible explanation is the existence of doublets, yet these tend to happen between specific pairs of cell types, which was not the most common case.

Overall, despite some losses in the resolution of cell groupings, the per-tissue clustering step of *CellTypist* maintains much of the cell type information existing in the original data.

3.2.2 Combining cell clusters across tissues using tissue-specific classifiers

In order to obtain a global reference for cell type classification, cell identity should be harmonised between all surveyed tissues. To achieve this, the clusters obtained at the end of the previous step are used as target labels predicted from gene expression using logistic regression models for each tissue (Figure 3.2A, left). Model characteristics and training parameters are detailed in Section 3.2.3. These models are then used to classify the complete datasets, thus obtaining assignment probabilities to the clusters in every tissue for each cell. These probabilities are further averaged per cluster, so that we obtain a mean probability of every per-tissue cluster corresponding to all others.

The merging of clusters is dependent on two parameters. The first (thr1) is defined as a threshold for the dot product of the mean probabilities of two clusters, above

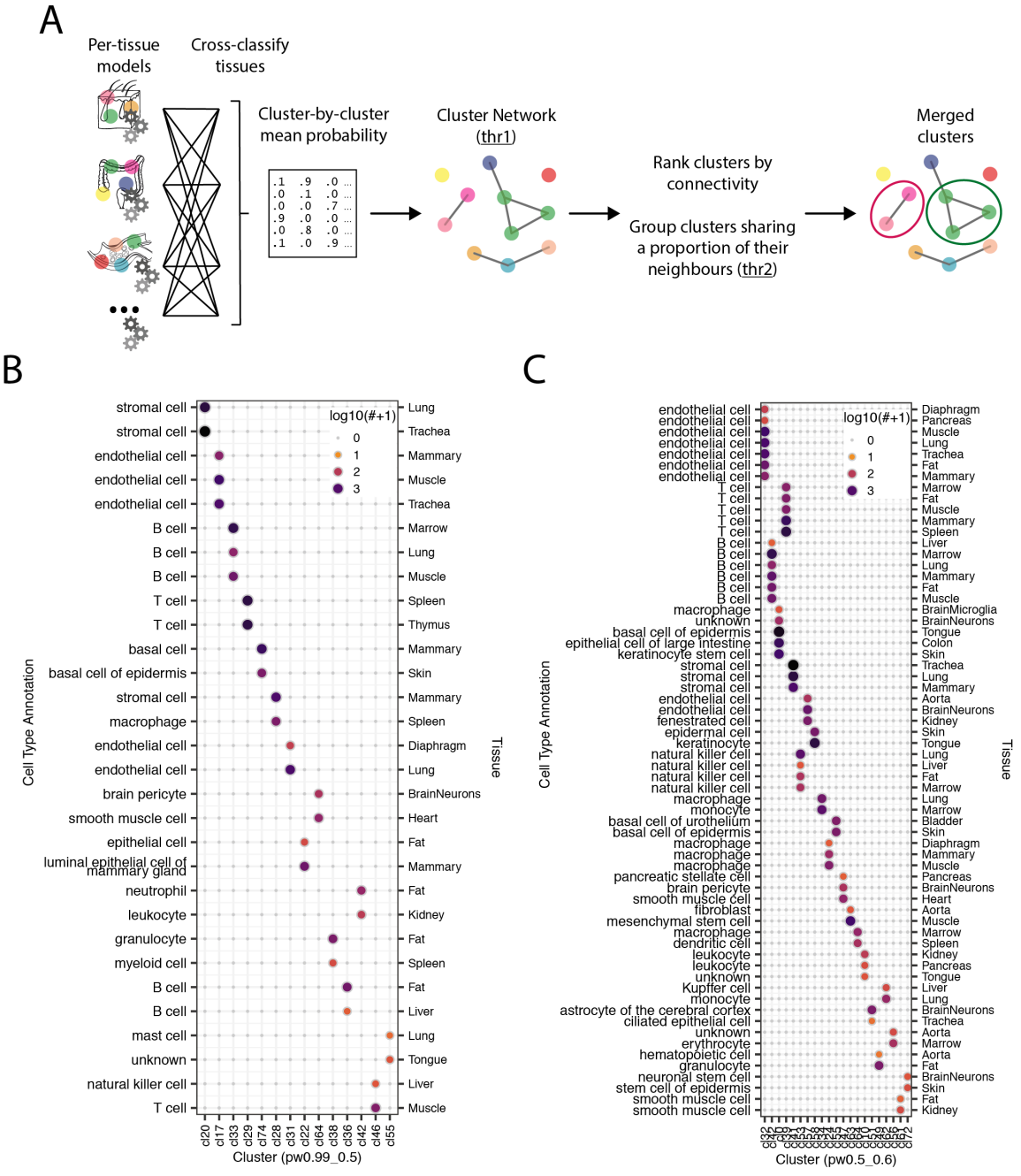


Fig. 3.2: **Cross-tissue matching of cell types**
(A) For each tissue, a logistic regression model is trained, and used obtain a classification probability of all tissues. Clusters are then linked depending on the mean probabilities of one cluster matching another (*thr1*). Clusters are ranked on connectivity, and grouped with neighbouring clusters that share a proportion of its neighbours (*thr2*). (Continued on the following page.)

Fig. 3.2: (continued) **(B and C)** Merging of cell types (x-axis) using the method in (A) for models trained on known cell type labels (left y-axis). **(B)** shows the top parameter combination (thr1 = 0.99, thr2 = 0.5) based on split-join distance; **(C)** shows the combination that came in third (thr1 = 0.5, thr2 = 0.6) and which resulted in increased merging.

which they are considered similar (i.e. "connected"). Based on this we can obtain a network with the connections between all clusters (Figure 3.2A, middle). This network serves as the base to define the cluster groupings. Clusters are then ranked based on their degree (i.e. the number of clusters they connect to) and grouped with their neighbours that share at least a defined percentage of their neighbours (thr2) (Figure 3.2A, right). The clusters merged are removed from the ranking, and the condition is iteratively applied until it has been tested for all elements. Lastly, the solutions for all thresholds are ranked based on their improvement over the per-tissue labels as measured by the split-join distance (detailed below).

To assess how this algorithm performed in a situation with known labels, it was tested using the annotated cell types for each tissue instead of clusters. After ranking the solutions given by the different parameter combinations tested (combinations identical to Figure 3.3B), we inspected the top parameter combination (Figure 3.2B, thr1 = 0.99 and thr2 = 0.5), as well as the third, which presented the lowest combined cluster number (Figure 3.2C, thr1 = 0.5 and thr2 = 0.6). Most clusters resulting from the merging workflow, in both solutions, combined cell types annotated with the same name in different tissues. This is particularly evident for endothelial cells, B cells, and T cells. While the score for the merging presented in panel C was not as good as that for panel B, it is immediately apparent that the more extensive merging still conserves most of the correct labeling, even grouping together identical cell types that are left separate in the first solution. This indicates that there can be a range of approximately correct parameter combinations, and hints at the tissue specificity of certain widespread cell types. Taking as an example endothelial cells, the first combination leaves lung and diaphragm separate from the remaining tissues.

This merging was then performed on the tissue clusters obtained from the first section of the pipeline. Both thresholds in the algorithm were tested with the values 0, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9, and 0.99. With the increase of both parameters, we observe an increase in the total number of clusters and a decrease in the fraction of merged clusters (Figure 3.3A). This trend is more evident for thr2, which is more directly involved in determining which clusters are grouped together. Results of all thr1-thr2 parameter combinations were then ranked based on

the split-join distance when comparing with known cell type labels, taking the original per-tissue clusters as a baseline. The parameter grid (Figure 3.3B) shows lower values for this ratio ("merged clusters" SJ distance/"per tissue clusters" SJ distance) at higher thr2 values, with the best combination (lowest ratio) at thr1 = 0.8, thr2 = 0.99. Examining this combination for how cell type labels were grouped revealed that, similarly to Figure 3.2B and C, many identical cell types had been grouped together

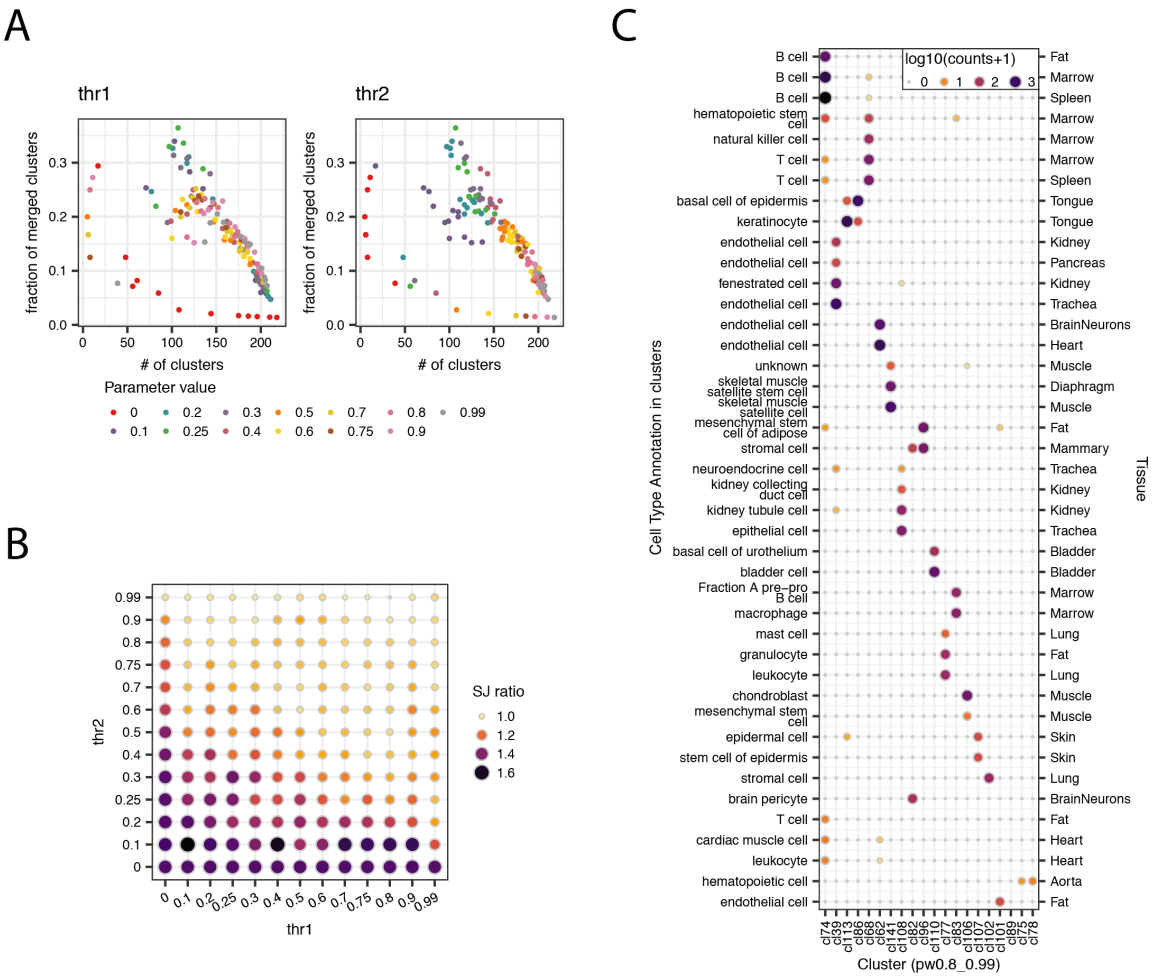


Fig. 3.3: **Evaluation of clusters matched across tissues**
(A) Change in number of total clusters and fraction of merged clusters with each threshold value (see Figure 3.2A for reference). Parameters resulting in a single cluster were not represented. **(B)** Parameter grid showing the variation of the ratio of split-join distance between merged clusters and cell type annotation, and per-tissue clusters and cell type annotation. **(C)** Grouping of cell types contained in per-tissue clusters (x-axis) using the top parameter combination (thr1 = 0.8, thr2 = 0.99) based on split-join distance.

(e.g. T cells; endothelial cells). Moreover, we can observe that cells with different names but similar functions are grouped together, as is the case of endothelial cells and fenestrated cells (an endothelial cell part of the renal glomerulus). In contrast, however, it could be observed that some cell types dispersed across more than one cluster. Even so, in cases where this happened (e.g. kidney tubule cell; mesenchymal stem cell of adipose), this dispersion tended to be minor, with a majority of cells from each of these annotated cell types coalescing in one cluster.

In sum, this demonstrates that this workflow is capable of merging cells with a similar transcriptome, and making cell identity across tissues uniform.

3.2.3 Generating updatable transparent-box models for cell type classification

Cell type classifiers have been descibed to achieve high performances even when based on simple models (Abdelaal et al., 2019; Köhler et al., 2019). The classifier used for *CellTypist* can be used to provide a fast and unbiased cell identity annotation of new datasets. *CellTypist*'s classifier is implemented in Python using scikit-learn (Pedregosa et al., 2011), and uses a logistic regression model with L2 regularization (Figure 3.4A). This allows the model to remain accurate, while still providing information about the contribution of all genes to determining the classification of each cell type. Training is done through mini-batch training using stochastic gradient descent (SGD). SGD is used since it makes the model more scalable, as it can converge without training over the whole dataset. It requires approximately one million data points to train, provided that all observations from all labels are passed to it. The models here presented will see the whole data a fixed number of times (epochs), to demonstrate their behaviour during training. The model encompassing all tissues was trained for 25 epochs, and the models trained on individual tissues (used in 3.2.2, Figure 3.2) were trained for 10 epochs. Additionally, SGD also allows for online training, meaning that if new data is obtained it can be easily incorporated into the model.

This methodology was first tested on the complete *Tabula Muris* dataset, training the model to predict the existing cell type annotations. The model converged after fewer than 500 iterations (each iteration corresponding to a batch of 1000 cells), and resulted in a prediction accuracy of 95% on the held-out test set (Figure 3.4B). Performance per class was assessed by calculating the F1 score, i.e. the harmonic mean between precision and recall for each class. A partial dependency between this score and the number of cells in a give class was observed for smaller groups (fewer

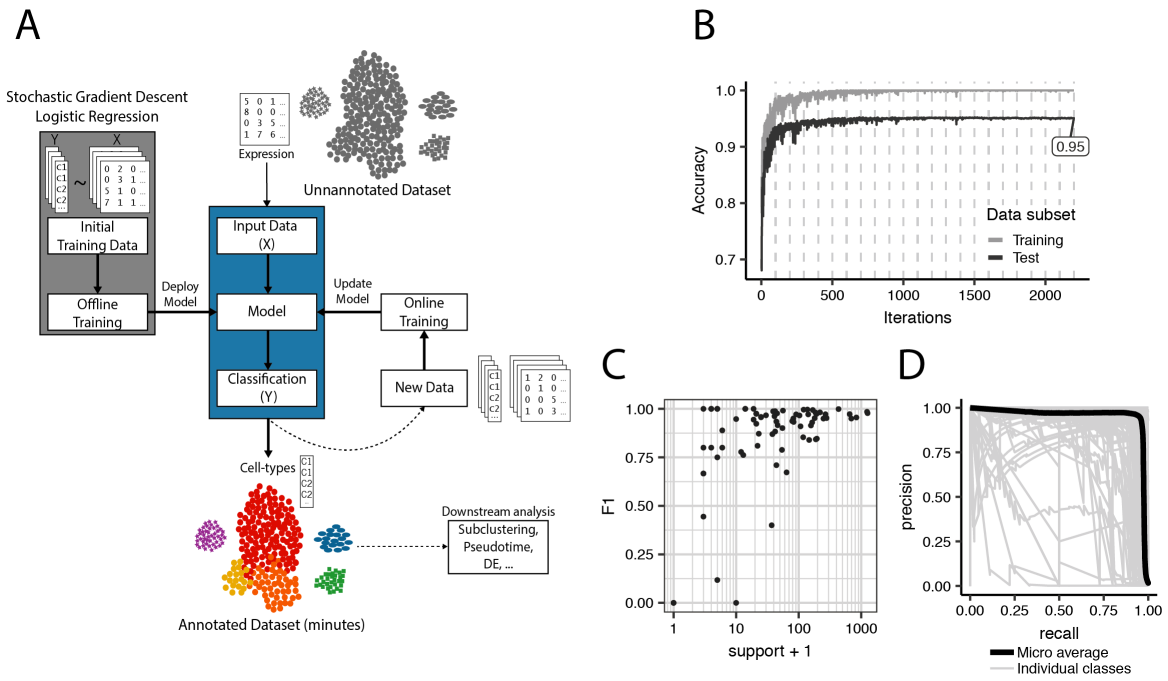


Fig. 3.4: **Model training outline and evaluation**
(A) Model training and usage outline. A logistic regression model is trained using stochastic gradient descent. When deployed, it can provide annotations for unlabelled data, which can the be further supplied back into the model to update it. **(B)** Accuracy during model fitting for training and held-out test data, to directly predict *Tabula Muris* cell type labels. Vertical dashed lines represent each training epoch. Terminal label indicated final accuracy for prediction in the test set. **(C)** F1-score for each cell type (black dots) as a function of class size (in log10 scale). **(D)** Precision-curves for each cell type (gray), and global micro average (black).

than 100 cells in the test set (10% of the total), Figure 3.4C, Tables B.1 and B.2). The strong predictive capability of the model can be further observed by plotting the precision-recall curves for each class (Figure 3.4D). While we again observe some classes to have a poorer performance, a micro-averaging of precision and recall of all classes (i.e. average precision and recall by calculating true positives, false positives and false negatives for each class) shows a very strong performance.

These results demonstrate the high performance of simple and intuitive logistic regression to train models capable of annotating data from various sources.

3.3 Results

3.3.1 Training *CellTypist* on the *Tabula Muris* dataset

After integrating scRNA-seq data across tissues, as described in Sections 3.2.1 and 3.2.2, the expression values can be used to unbiasedly predict cell identity by constructing a model in a similar fashion to that described in Section 3.2.3.

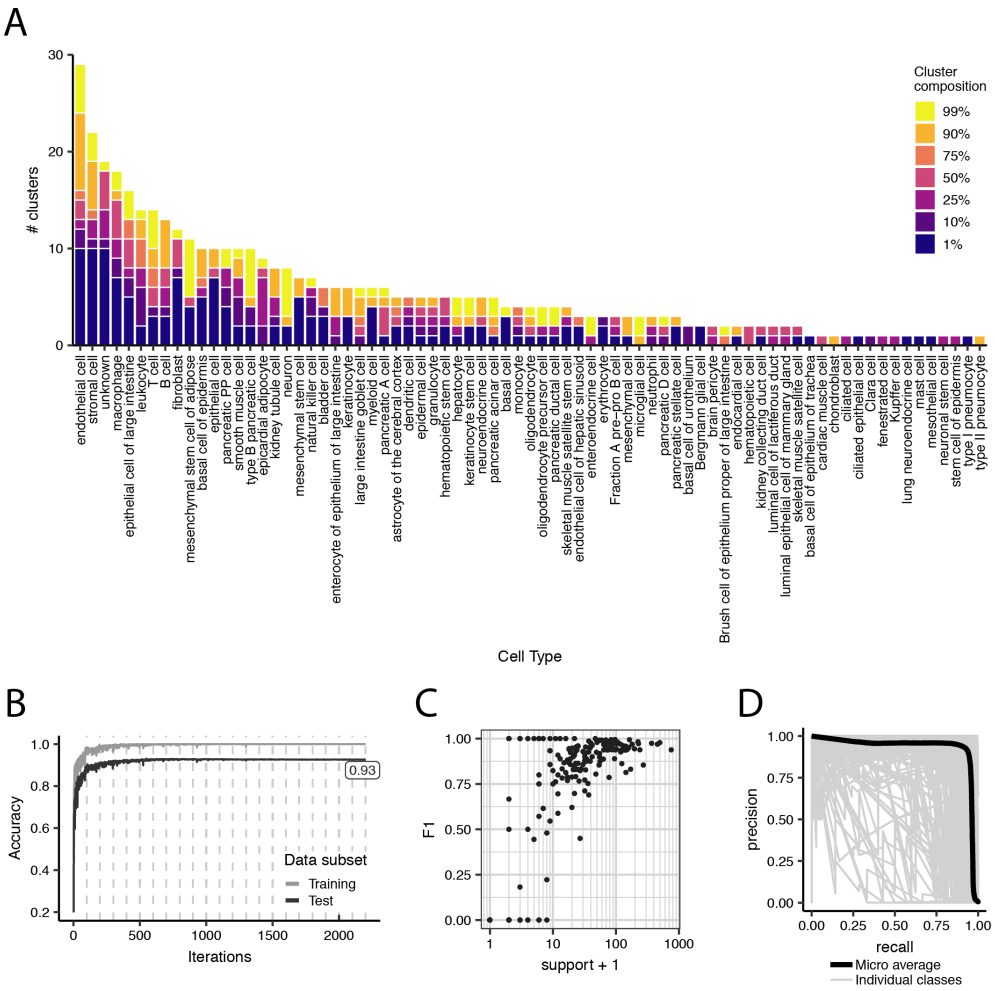


Fig. 3.5: **Evaluating model trained on cross-tissue integrated clusters**
(A) Abundance of annotated cell types in cross-tissue clusters. Colours represent clusters with at least x% of a given cell type. **(B)** Accuracy during model fitting for training and held-out test data, to predict cross-tissue integrated clusters. Vertical dashed lines represent each training epoch. Terminal label indicated final accuracy for prediction in the test set. **(C)** F1-score for each cluster label (black dots) as a function of class size (in log10 scale). **(D)** Precision-curves for each cluster (gray), and global micro average (black).

The model training framework was thus tested using the cluster labels resulting from the merging shown in Figure 3.3 (thr1 = 0.8, thr2 = 0.99). Clustering at the tissue level resulted in 222 clusters overall, compared with 139 cell type-tissue combinations. This increase was mainly registered in Pancreas (+20 clusters), Colon (+17), Fat (+11), Brain Neurons (+11) and Aorta (+9). All of these were clustered with a resolution of 0.1, and all were solely represented in Smart-seq2-derived data (Figure B.1). This hints at potential batch effects that were unaccounted for in the initial analysis step. Merging the clusters across tissues resulted in a final number of 198 clusters as the top result, compared to 75 unique annotated cell types, a difference that is mostly propagated for the initial large increase in the number of clusters. Nonetheless, those that were merged grouped in many cases cells with a similar phenotype in many cases (Figure 3.3).

Figure 3.5A examines the representation of annotated cell types across all clusters, showing that a large majority of cell types are in one or more clusters where they represent at least 90% of cells, indicating that although the number of clusters is elevated compared to expectations, most clusters are highly specific. Similarly to the cell type-based model, performance metrics globally show a fast convergence and high training accuracy (Figure 3.5B), as well as a high per-label precision and recall (Figure 3.5C, D), with most classes having an F1 score above 0.75. Classification performance can be seen broken down by cluster in Tables B.3, B.4, B.5, and B.6. These again reveal the poorer performance of lowly represented clusters, many of which originating from the overclustered tissues mentioned above.

3.3.2 Training *CellTypist* on a collection of human data

To obtain a global, cross-tissue perspective of human cell types, we obtained a broad representation of single-cell transcriptomes by collecting several publicly available scRNA-seq datasets (Table B.7). Information about tissue, scRNA-seq protocol, sampling method, and cell type annotation (when available) were obtained from the respective publications and data repositories, together with the gene expression matrices (Figure 3.6).

The 28 datasets collected include 21 tissues, mostly collected from adult biopsies (Figure 3.6A and 3.6B), and totalling close to 1.5 million cells. Various studies focus on haematopoietic-derived cells, and as such many of the sampled tissues are mostly composed of immune cells (Figure B.2). Most cells are obtained using the droplet-based "Chromium" instrument from 10x Genomics ("10X" in Figure 3.6C),

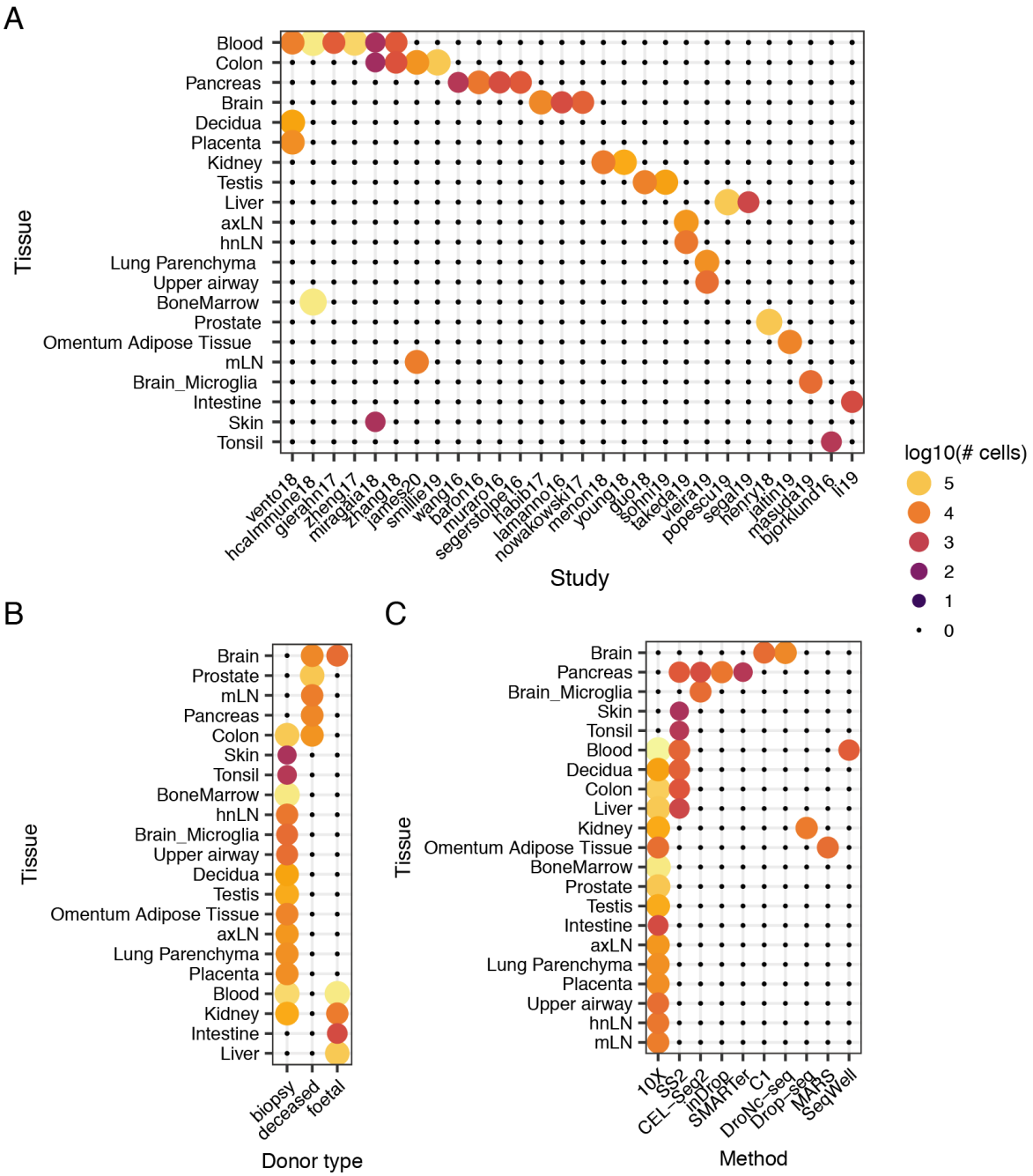


Fig. 3.6: **Cell numbers in the human dataset collection**
Number of cells, in log10 scale, collected from different tissues, and distributed by publication **(A)**, type of collection **(B)**, and scRNA-seq protocol **(C)**. hnLN - head and neck lymph nodes; axLN - axilary lymph nodes; mLN - mesenteric lymph nodes.

followed by the plate-based, full-length Smart-seq2 ("SS2"). Despite this imbalance in usage of different technologies, it is in agreement with what has been reported

in an exhaustive curated reference of single-cell sequencing datasets (Svensson and Beltrame, 2019).

Single-cell RNA-seq expression data was collected for the publications listed in Table B.7, together with cell type annotations when these were available. Information about tissue, donor type and scRNA-seq protocol were obtained from the publications.In most cases, count data was available together with the raw sequencing reads in the chosen repository. In other cases, the expression matrices deposited included log normalised data. This means that the data was normalised by the total number of reads/UMI of each cell, often followed by multiplication by a specific scaling factor (usually 10000), and finally log scaled, adding 1 to account for the zeroes present. For these datasets, data was reconverted to counts following an approach similar to that explained in http://www.nxn.se/valent/2018/10/25/unscaling-scaled-counts-in-scrna-seq-data. Briefly, given the scaling factor S, representing the second most abundant value for each cell, and x for each expression value, unscaled data U was obtained by applying Formula 4.1, followed by rounding to the nearest unit to remove floating point inaccuracies.

$$U = \frac{e^x - 1}{S} \tag{3.1}$$

Raw count matrices were then compiled together, guaranteeing as much correspondence as possible between the diverse gene references used. All gene identifiers were mapped to the corresponding HGNC gene names, and all unique identifiers were kept. This was done to maintain the integrity of each dataset, as well as facilitate data collection and incorporation.

The *CellTypist* pipeline was then applied to the complete human dataset, with parameter optimisation as described in the previous Sections. Data from the same tissues was integrated and clustered using the Leiden algorithm (Traag et al., 2019) at several resolutions. For tissues with cell type annotations, resolution was optimised using the split-join distance (Dongen, 2000) between clusters and cell type annotation and constrained to a number of clusters at least as large as the number of cell type annotations in the largest collected dataset (Figure 3.7A, see Section 3.2.1). This led to a total of 641 clusters in all tissues.

Following clustering, per tissue logistic regression models were trained, running for 10 epochs of a maximum of 100 iterations each. These models were used to run the cross-tissue cluster merging pipeline (Section 3.2.2), and a combination of parameters was chosen based on the ratio of split-join distances (merged vs annotated cell types

over per tissue vs annotated cell types) (Figure 3.7B, Figure B.3A,B), resulting in the choice of thr1 = 0.99 and thr2 = 0.8 (627 clusters). Additionally, three other combinations were chosen for comparison: thr1 = 0.4 and thr2 = 0.99 (607 clusters), the combination with the top split-join ratio when only considering merged clusters (Figure B.3C, Figure B.5A-B); thr1 = 0.25 and thr2 = 0.25 (420 clusters),

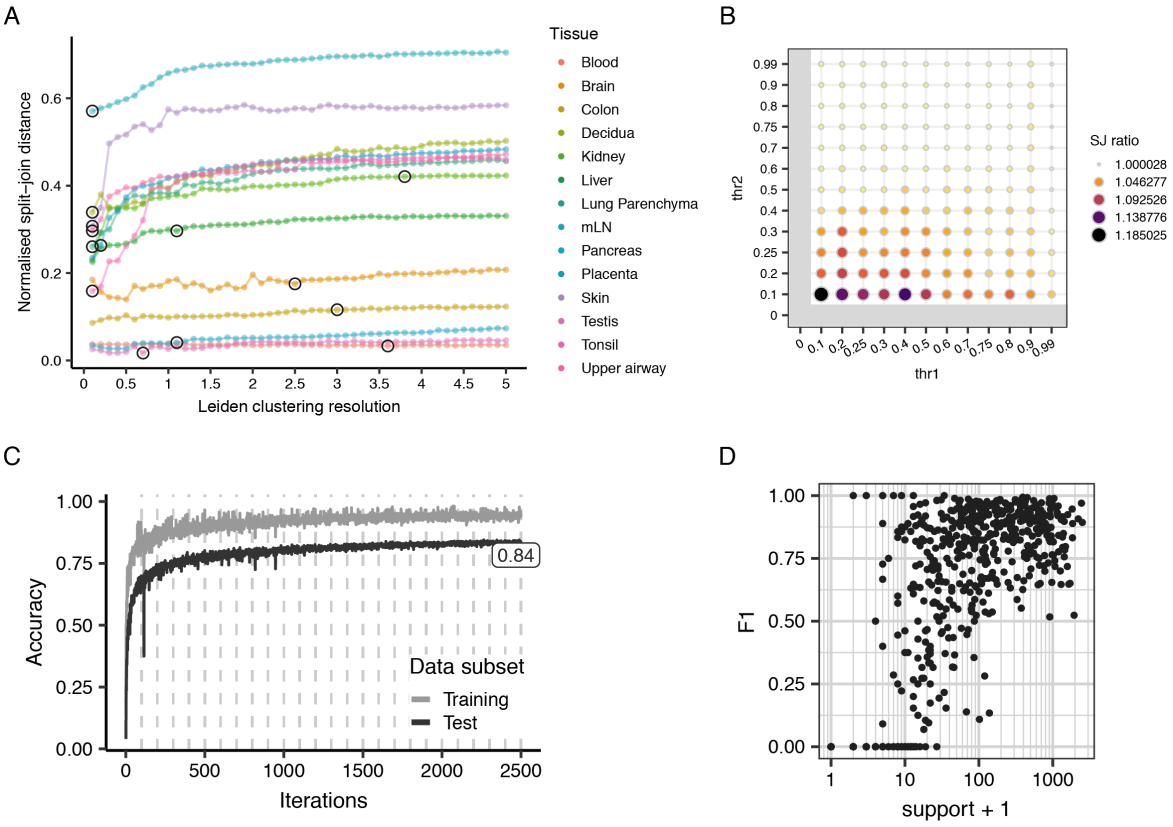


Fig. 3.7: **Running *CellTypist* on a human scRNA-seq data collection**
(A) Per-tissue cluster optimisation, choosing the resolution that approximates existing cell type annotations. Similarity is measured with normalised split-join distance, and constrained to solutions with a number of clusters of at least as many as existing annotations in the largest collected dataset. Selected values are indicated with a black circle. **(B)** Grid of parameters tested for cross-tissue cluster merging, showing the variation of the ratio of split-join distance between merged clusters and cell type annotation, and per-tissue clusters and cell type annotation (colour and size of points). **(C)** Accuracy during model fitting for training and held-out test data, to predict cross-tissue integrated clusters obtained using thr1 = 0.99 and thr2 = 0.8 as parameters for *CellTypist* (optimal value in (B)). Vertical dashed lines represent each training epoch. Terminal label indicated final accuracy for prediction in the test set. **(D)** F1-score for each cluster label (black dots) as a function of class size (in log10 scale).

one of the combinations with the highest fraction of merged clusters (Figure B.3B, Figure B.5C-D); thr1 = 0.1 and thr2 = 0.1 (218 clusters), the combination with the highest fraction of merged clusters, as well as highest split-join fraction (Figure B.3B, Figure B.5E-F).

An example of a tissue (pancreas) with consistently annotated cell types across datasets can be seen in Figure B.4 with the merged clusters in the thr1 = 0.99 and thr2 = 0.8 model. We can appreciate that the pipeline, similarly to Figure 3.3C, successfully merged various similarly annotated cell types (alpha, beta, acinar, ductal, delta, gamma, epsilon, endothelial), albeit with some small "contamination" by other cell types. Other however were not so well separated, as is the case of the immune cells (t_cell, mast, MHC class II), which were grouped together.

The cell groupings obtained were used to train a logistic regression model using Stochastic Gradient Descent (Section 3.2.3). Training was done for 25 epochs of a maximum of 100 iterations each, where in each iteration 1000 cells were seen by the model. 90% of the total data was used as a training set, and the remaining as a left out test set that was tested at every iteration (Figure 3.7C, Figure B.5). The model had a classification accuracy of 84% on left-out test data (Figure 3.7C), and the F1 statistic calculated for each label was in most cases above 0.75(Figure 3.7D, Tables B.8 to B.19), meaning elevated precision and recall in the model, especially for clusters with more than 100 cells, as previously shown (Figure 3.4C, Figure 3.5C).

Three additional models, trained using sets of clusters derived using different parameters, were also examined (top markers for largest clusters from each model are listed in Supplementary Table B.20). These were thr1 = 0.4 and thr2 = 0.99 (Figure B.5A-B), thr1 = 0.25 and thr2 = 0.25 (Figure B.5C-D), and thr1 = 0.1 and thr2 = 0.1 (Figure B.5E-F). These models show a lower performance, in particular the latter (thr1 = 0.1 and thr2 = 0.1), with a test classification accuracy of 73%. This may be due to the excessive merging of clusters within and across tissues, thus leading to hybrid, undetermined groups of cells (Figure C.4). Other models are more conservative in this regard, and show a better classification performance. The model using clusters obtained with thr1 = 0.25 and thr2 = 0.25 still has noticeably worse values for the split-join distance, yet also represents a more condensed cell type reference (420 clusters), without sacrificing accuracy (83%). Lastly, the model with the parameters thr1 = 0.4 and thr2 = 0.99, in particular, is the one showing the greatest improvement in matching annotated cell types after cross-tissue merging, thus representing another possible reference model.

In sum, the collected datasets allow for the training of the *CellTypist* pipeline and construct a fully interpretable human cell type reference.

3.4 Discussion

CellTypist has been designed as a way of systematising cell identity from expression data, and use it directly for automatic annotation. The pipeline has been designed keeping scalability in mind, fully aware that the first model represents an initial release that will be continuously updated. It is expected that the increase in data sources and available expert annotations will greatly improve the usability of the framework going forwards.

The construction of this pipeline is also subject to evolution. It has been developed with the ability to include unannotated data in a cell type reference. Existing cell type annotation is highly informative when deposited together with the expression data or the accompanying publication, although this is not always the case. Even so, the vast majority of scRNA-seq analysis pipelines rely either on Leiden (Traag et al., 2019) or Louvain (Blondel et al., 2008) clustering, which are used in the per-tissue processing step and thus results in a considerable approximation between known and new labels (Figure 3.1B). Even though the final, merged labels are to be manually curated and named, existing cell type annotations can also be made available to the end user, adding another layer of validation to the results.

The results presented here demonstrate that integration using the pipeline can correctly merge similar cell types (Figures 3.2B,C; Figure 3.3C; Figure B.4C). However, these results are not perfect. The fact that, in some instances, T cells share clusters with natural killer cells, highlights that the method does not yet achieve perfect cell type separation. These two cell types have similar transcriptomes, which explains why in some reduced representations used for clustering they might appear very close. Nonetheless they are easily distinguishable by the expression of a small set of markers, and could be efficiently distinguished by a logistic regression model using curated labels (Figure 3.4C and D). This mixing between different cell types results from both integration stages of the pipeline. Figures 3.3C and B.4 show that the merged clusters are composed of more than one annotated cell type, although generally containing a majority of a specific cell type, which results from the non-exact overlap between per-tissue clusters and annotated cell types. Additionally, misgrouping caused by the cross-tissue matching step (Figure 3.2B) can occur, where similar cell types can

be incorrectly merged (e.g. natural killer cells and T cells; smooth muscle cell and epithelial cells).

These inaccuracies can be mitigated in various ways. The resolution bottleneck introduced by the first, per-tissue integration step can be potentially improved. For example, one possibility is to adopt a more curated approach after clustering every tissue, although this would require significantly more human input. Another option would be to rely more on data with existing annotations. One way to achieve this is by using a label propagation method that, within each integrated tissue, passes existing labels to unannotated data (Barkas et al., 2019). Alternatively, the method can instead iteratively apply the algorithm used to integrate the clusters between tissues (Figure 3.2A). In the first step this would be applied for each dataset collected to merge all existing data for each tissue, and relying on existing annotations. A second step would then be applied between tissues as shown (Figure 3.2). This guarantees that any known heterogeneity in the collected data is preserved and propagated into the final annotation, but has the disadvantage that any novel populations that would be detected by data integration can be lost, and requires the cell type labels to be thoroughly verified *a priori*. Integration can also be improved by a cross-tissue batch alignment approach (e.g. using MNN (Haghverdi et al., 2018) or BBKNN (Polański et al., 2019)), which could potentially help with the proper overlap between cell populations, yet can be difficult to apply at scale. It is expected that improving the integration steps of the pipeline will ultimately lead to more confident matches across tissues, since at the moment the pipeline is very conservative when establishing these relationships. Lastly, an ensemble approach can also be taken by combining all the per-tissue models into a single classifier. This has the disadvantage that the pipeline will not immediately identify relationships between cell types in different tissues, yet could potentially, from the same model, match the most similar cell population and tissue from the same model.

This Chapter also demonstrated the viability of logistic regression as a methodology for cell type classification using a broad atlas as a reference (Figures 3.4B-D and 3.5B-D). This is in line with previous reports (Abdelaal et al., 2019; Köhler et al., 2019), showing that cell identity classification is not improved by the use of deep learning methods, and can be accurately performed using simpler machine learning frameworks. The results also demonstrate that this method is robust enough to accommodate clusters with some mixture of cell types (which results in lower phenotypic resolution) (Figure 3.5), and even from a broad variety of tissues and protocols (Figure 3.7). However, it is important to highlight the potential biases that

can arise from the collected data, since the representation of cell types in different tissues might be uneven. Differential representation of cell populations across tissues can potentially result in a bias learned by the model. As an example, if a cell type has a tissue-specific signature (that is not present in other cells from that same tissue), and this cell type is more abundantly profiled from that tissue, the gene signature learned by the model may reflect, at least partially, the tissue-specific signature rather than the desired cross-tissue phenotype. While this is difficult to consider for all cell types (due to likely tissue-specific heterogeneity and lowly abundant populations), the use of down-sampling (Hie et al., 2019b; Wong et al., 2016b) before training or the application of a model ensemble approach could mitigate this. Furthermore, imbalances in the data can also result in differential representation of cell populations across tissues. This is the case with the human data collection, where there is a preponderance of immune cells, with some organs profiled only with respect to this cell compartment (Figure B.2). This justified the need for an updatable model, in order to add data to incompletely profiled tissues.

The implementation of *CellTypist* is also explicitly constructed such that the resulting model can be easily updated. This is due to the implementation using stochastic gradient descent, which allows for easy and direct updates to the model by running more learning iterations on novel data. This is important to maintain the reference up to date. However, it can only be done by classifying new data according to the existing labels. If the new datasets collected include cell types that are not represented in *CellTypist*, then the full pipeline needs to be ran anew. Nonetheless, this allows for the database to have fast minor releases to maintain it up to date with the latest dataset publications, as well as less frequent major releases that more thoroughly integrate these datasets and revise the annotation database. *CellTypist* will be available with a web interface, allowing for classifications to be ran via a web server, or, alternatively, download the models to test locally. It will include a database characterising the cell type labels present in the model. While different organisms will have different models, many of the cell types described are predicted to be present in multiple species, and can in future updates have their cross-species similarities defined and reflected in *CellTypist*'s accompanying cell type compendium.

Overall, this chapter has demonstrated that *CellTypist* can organise a valuable resource for cell type annotation. Furthermore, this resource can be readily interpreted from inspection of gene coefficients for each label. In the next chapter, a practical application and interpretation of *CellTypist* will be presented.

Chapter 4

Application and biological insights of the *CellTypist* model

The identity of a cell can be defined by the genes it expresses. Knowing these gene sets is what helps us to identify cell types when analysing scRNA-seq data, yet this manual identification often requires a vast domain-specific knowledge, and thus interpretable models that either rely or identify these genes can be useful to make cell type classification automatic. Furthermore, an increasing number of studies has applied scRNA-seq to profile various body locations and describe the cells that make up a tissue, in the steady-state or disease. A smaller number of studies have focused on the differences between the cell types detected in different tissues (Miragaia et al., 2019; Scott et al., 2018). However, we don't yet know how variable the transcriptome of most cell types is between tissues, and how much of a tissue's transcriptomic identity is reflected in each cell type.

This Chapter follows from the construction of the *CellTypist* models in the previous Chapter, and explores its application for automatic cell type classification and interpretability with regards to cell identity and cross-tissue relationships. The present Chapter will reveal the type of genes important in defining cellular phenotypes across tissues, and outline how tissue gene expression signatures relate different anatomical regions.

The analyses here performed are based on the methodology outlined in Chapter 3. Supplementary figures and tables are included in Appendix C.

4.1 Introduction

Recent developments in single-cell sequencing have enabled unbiased and high-throughput assessment of cell types through transcriptomic profiling (Svensson et al., 2018). A few individual works have aimed at profiling cell types across most tissues of an organism (Fincher et al., 2018; Han et al., 2018; Plass et al., 2018; Various, 2018). Other more complex and detailed cellular census have been done for individual tissues, and large consortia have been established to aggregate some of these datasets and establish guidelines and collaborations to identify all cell types across an organism (Regev et al., 2017).

The definition of cell type is, like many biological terms, a working definition. Cells have been classified based on different aspects of their morphology, molecular phenotype, or function. Historically, this knowledge of cell identity has been restricted to specific fields (e.g. immunology, neuroscience), hindering the development of an integrative, systemic perspective of cell types in the body. Single-cell RNA-seq technologies (scRNA-seq) are now challenging this perspective, since they allow for an unbiased profiling of cell identity through the transcriptome. As scRNA-seq data acquisition grows (Svensson et al., 2018), so does our understanding of the cellular make up of the profiled tissues. The Human Cell Atlas Consortium has defined as one of its goals to develop a cellular taxonomy (Regev et al., 2017), which is necessarily harmonised across tissues. Nonetheless, a unified, transcriptome-driven perspective of cell identity is still lacking.

The molecular basis for the relationships between tissues were initially probed by high-throughput methods; first microarrays (Enard et al., 2002), and later with RNA-seq (Barbosa-Morais et al., 2012; Brawand et al., 2011; Mortazavi et al., 2008). More recent studies are now linking this transcriptome cross-tissue variability with genome variants (Consortium, 2015; GTEx Consortium, 2017), unravelling the regulatory determinants behind tissue biology. Further analysis have delved into the importance of transcription factors for tissue identity (Sonawane et al., 2017), revealing that tissue specificity lies not only in these molecules but mostly on the tissue-specific regulatory roles they play, while also showing that transcription factors are less likely to be identified as tissue-specific than other genes. An integrated predictive model of cell identity should be able to reveal patterns relating tissues through cell identity relationships, as well as offer a broad perspective of the genes determining cell types.

Here we will expand on the pipeline developed in Chapter 3, testing *CellTypist* for automatic annotation of scRNA-seq data, to probe cellular identity in primary cells

across body locations. Testing the model trained on human data on an independent dataset reveals an elevated accuracy for cell types and tissues represented in the reference, as well as informative approximations for cell types not yet included in *CellTypist*.

Beyond classification, *CellTypist* can also be dissected to unravel aspects of cell type and tissue biology. The integration pipeline can recapitulate known tissue associations, caused either by comparable cell sampling (e.g. tissues solely profiled for immune cells), or by functional similarity. These are evident both at the tissue integration stage, as well as in the top genes learned by the model that define cell groupings in each tissue. These genes are further examined for patterns in cell identity definition, revealing a global pattern for genes coding for functional effector molecules (i.e. receptors and secreted proteins) to be more pivotal in defining cell identity than others involved in genomic regulation. Finally, we discuss the potential uses and implementation of a scRNA-seq-derived cell type reference.

4.2 Results

4.2.1 CellTypist as an operational reference for annotation

The operational goal of *CellTypist* is to be used as an automatic classification framework for scRNA-seq data. Data integrated through the pipeline can be used as an unbiased model of cell identity to predict cell type labels in unannotated data.

The data generated in (Madissoon et al., 2019) was used to test the classification performance of *CellTypist* with the compiled human data. This dataset was chosen because it includes three distinct tissues - lung, oesophagus, and spleen. Of the three tissues, lung is the only represented in the collected datasets (yet not contributed from the same sample), although many of the cells collected from spleen (mostly immune cells) are present in the model, contributed by different tissues. More than 200.000 cells were collected from these three tissues, with various cell populations manually identified.

An overview of the classification results, projected in UMAP (McInnes et al., 2018) (Figure 4.1A, Figure C.9A, Figure C.10A), shows a similarity between the individual labelling of different clusters. The increased noise in *CellTypist*'s annotations are likely due to the large number of categories it includes. Despite this, most model labels are highly specific, being attributed almost entirely to a single original cell type annotation (Figure 4.1B, Figure C.9B, Figure C.10B). While the opposite is not

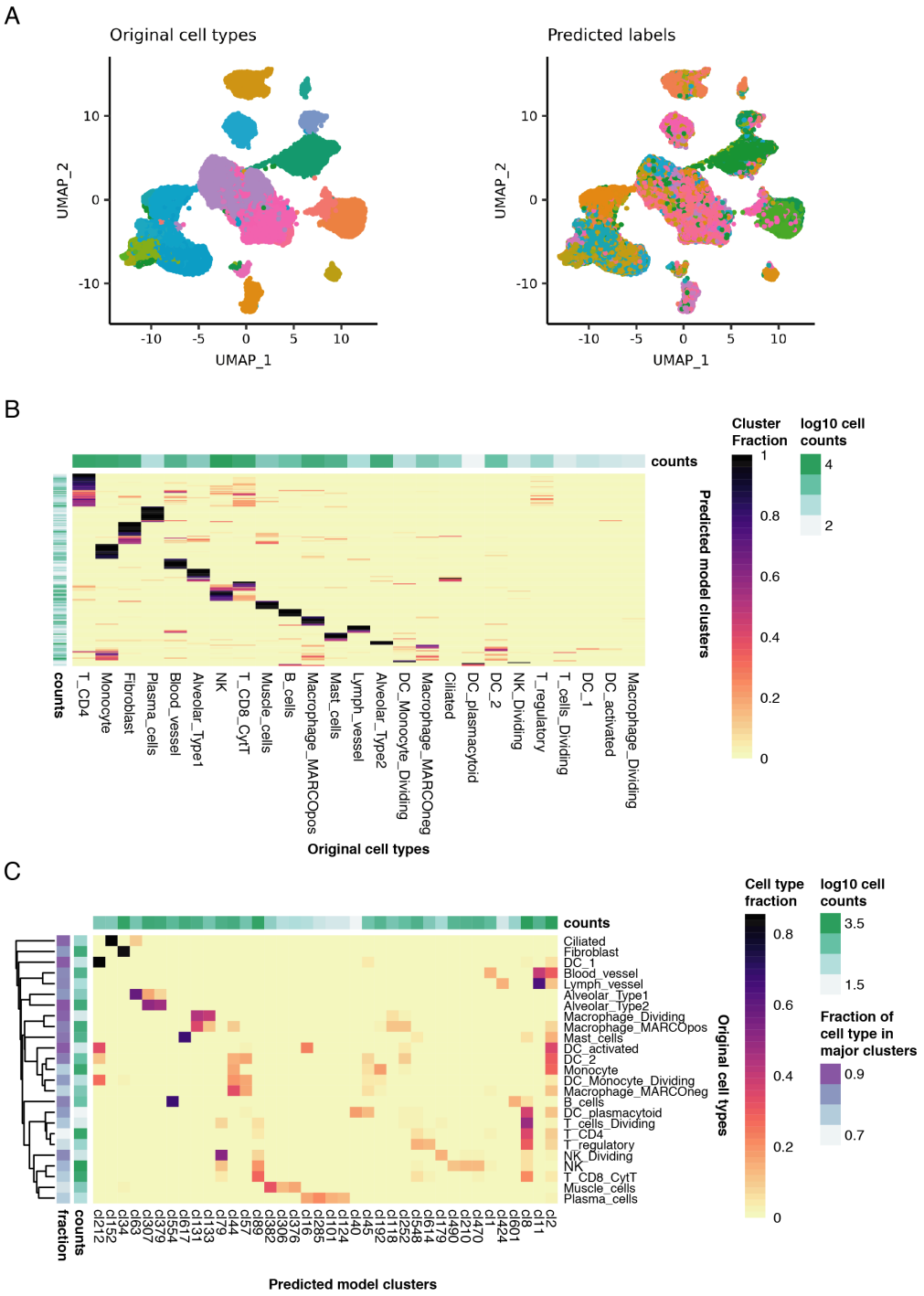


Fig. 4.1: ***CellTypist* predictions for lung data from (Madissoon et al., 2019)** **(A)** UMAP projections coloured by the original cell type annotations (left) and those predicted by *CellTypist* (right) using thr1 = 0.99 and thr2 = 0.8. **(B)** Proportion of clusters (rows) matching each annotated cell type (columns). **(C)** Proportion of annotated cell types (rows) included in each cluster (columns). Only clusters including at least 10% of a given cell type were included.

true (i.e. one cell type annotation can correspond to more than one cluster), it is nonetheless evident that each cell type is dominated by one or very few clusters (Figure 4.1C, Figure C.9C, Figure C.10C). Furthermore, even when excluding clusters including less than 10% of cells from each annotated cell type (as is the case in the heatmap in Figure 4.1C), the remaining clusters still include 70-90% of cells (purple sidebar in Figure 4.1C).

A downside of validating *CellTypist* with independent data is that the comparison can not be directly assessed, since the existing annotations for this dataset do not match those used by *CellTypist*, which compiles a variety of nomenclatures used in each specific publication from where the data was obtained. Nonetheless, this can be circumvented by manual inspection of existing labels, as well as matching the dataset annotations with the model clusters to approximate a gold standard.

A more careful look at the annotations present in the clusters that matched each original cell type in lung reveals the accuracy of the model. Type 2 alveolar cells matched clusters only containing that same annotation, whereas clusters in alveolar type 1 included type 1, and type 2, as well as secretory cells. Ciliated cells and fibrolasts mostly matched a single cluster each, in both cases composed of the exact same annotation. Cells annotated as "Lymph_vessels" and "Blood_vessels" both matched cl11 (containing "endothelium" and "lymphatic" cells), with the first also matching lymphatic endothelial cells from the axillary lymph node. T cell annotations were mostly assigned to cluster cl8, which includes a mix of CD4 and CD8 cells. In addition, T regulatory cells also matched cluster cl614, which includes activated T cells and Tregs. NK cells also matched a cluster with CD8 T cell annotation, but included two others containing mostly NK cells from other tissues. Lung cells that are derived from the myeloid lineage (Macrophages, Monocytes, Dendritic cells) all matched clusters mostly composed of these same annotations, albeit with some mix between them, which again demonstrates some of the difficulty that exists in separating these cell types.

For a more quantitative assessment of *CellTypist*'s accuracy, an identical cell type nomenclature would have to exist between the model and the validation dataset. While one could opt for renaming *CellTypist*'s labels in accordance with those present in the model, two arguments invalidate this approach. First, converging into the exact same labels could be misleading, since different methodologies would be utilised for labelling the validation dataset and the model clusters - the latter would rely on the model coefficients, as well as existing annotation from the original datasets. Second, this would not take into account differences in resolution between the model

Fig. 4.2: **Classification accuracy for the (Madissoon et al., 2019) dataset** **(A)** Variation in mean accuracy with the maximum number of cell types from each tissue allowed to be matched to a cluster. Weighted mean takes the number of cells into account. **(B)** Variation in mean accuracy with the maximum number of cell types from each tissue allowed to be matched to a cluster, stratified by tissue. **(C)** Classification accuracy for each cell type from each tissue (colour), as a function of the entropy calculated for the predicted cluster representation of each cell type, normalised by its representation.

and the data, and would penalise its lack of specificity. An example of this is the situation described in the previous paragraph for the dataset's "Lymph_vessels" and "Blood_vessels" labels, which both match a clearly general endothelial cluster, and thus opting for one of these labels would wrongly penalise the other.

Instead, correspondence between dataset cell types and model clusters was independently determined by assessing the enrichment for cell type markers in the top 500 coefficients for each model label using GSEA (see Methods Section 4.4.2). This was done in an attempt to approximate the annotations based on marker gene expression, the most commonly used methodology. All tissue/cell type combinations were tested together for enrichment, filtered for significance (q-value<=0.05) and positive enrichment scores, and ranked by the latter. Cluster-cell type correspondence was assessed per tissue, with a variable number of corresponding top cell types accepted (Figure 4.2A and B). Accuracy was then calculated for each cell type, based on whether each cell's cluster assignment by the model had was enriched for the same cell type originally labelling that cell. As expected, inclusion of more cell types to match each cluster led to increasing accuracy (Figure 4.2A).

This accuracy was different between tissues, with lung as the best scoring, followed by Spleen and the Oesophagus (Figure 4.2B). This is in line with the composition of the data that underlies *CellTypist*: Lung has a high accuracy since this tissue and most of its cell types are represented; Spleen also has elevated scores since it is mostly composed of immune cells, which are highly abundant in the model coming from Blood, Bone Marrow, and other tissues; Oesophagus presents a lower score due to the fact that the sample mostly includes different types of specialised epithelial cells, which are absent or underrepresented in the model.

The accuracy per cell type and tissue was then examined (Figure 4.2C), allowing for up to 5 cell types to correspond to a cluster, the value at which accuracy stabilises. These assignments can be found in Supplementary Table C.1 to Supplementary Table C.10. A value greater than 1 also has the advantage of better reflecting the many-to-many relationships that exist between model clusters and manual cell type annotations. It should also be noted that this implies that the model cluster can represent a lower resolution or broader cell type identity in many cases. For example, macrophages and other myeloid cells are enriched within the same cluster despite consisting of many (sub)types of cells (for a concrete example, see cl252 in Supplementary Table C.3).

Figure 4.2C shows an increased accuracy for most immune cell types, as well as lung-derived cells. The lowly-performing immune cells from the spleen originate

from rarer populations, which explains the higher weighted mean accuracy, and reveals resolution limitations in the model. Conversely, most of the top-performing cells from the Oesophagus come from immune cell populations with a low level of specificity (as illustrated by the"NK_T_CD8_Cytotoxic" label), which also make up a small percentage of the total cells recovered from this organ. There is a modest negative correlation between the accuracy for each cell type and how many clusters each cell type is distributed across (normalised by log10(number of cells), Spearman Correlation = -0.33, p-value<0.01). Lastly, Figure 4.2 shows that accuracy from cell types and tissues represented in the models will be in the range of 0.4 to 0.85, and rare or absent cell populations will be between 0.2 and 0.5.

The other models resulting from different parameters were also briefly examined. Despite the differences in number of clusters, all models show a similar specificity for the assigned clusters (Figure C.12). However, both models with fewer clusters (thr1 = 0.25, thr2 = 0.25; and thr1 = 0.1, thr2 = 0.1) both show less unique matching of original cell types to clusters (Figure C.13), with most of them matching the same larger clusters, which is likely an artifact of excessive merging within and across tissues (Figure 3.2A).

Globally, it has been demonstrated that *CellTypist* can be successfully used to annotate datasets with a broad diversity of cell types, and future improvements to the pipeline are likely to make it more precise in attributing cell identity.

4.2.2 Matching cell identity across tissues

The number of clusters detected in each tissue are independent of the number of datasets (Spearman Rank Correlation = -0.01, p-value for null hypothesis of "$\rho=0$" = 0.9344), although moderately correlated with the number of cells present in each tissue (Spearman Rank Correlation = 0.52, p-value for null hypothesis of "$\rho=0$" = 0.01497) (Figure C.2). The subsequent cluster merging step draws a map of cell identity relationships across tissues. Examining this map can reveal higher order relationships between the tissues present in the global dataset. Thus, the per-tissue classification probabilities used to construct the cluster matching graph (Figure 3.2A) were used to calculate the mean probability of cells from a per-tissue (non-merged) cluster matching the clusters of all tissues. The resulting tissue-by-cluster mean probability matrix is represented in the clustered heatmap of Figure 4.3A. This plot shows that about a third of all clusters have an average high confidence assignment

across tissues (bottom of the heatmap), with the remaining two-thirds having much lower per-tissue mean probabilities.

The clustering of these values reveals a stark division between tissues whose immune compartment was predominantly profiled (left major branch of dendrogramme), and those with a more global or non-immune profiling (right branch). This is highlighted by the per tissue mean expression of *PTPRC*, the gene encoding for the CD45 receptor, which is exclusively expressed in immune cells (Altin and Sloan, 1997) (Figure B.2). Expression of *EPCAM* - an epithelial cell marker - and *CD34* - an endothelial cell marker - further illustrate this division, being most expressed in tissues in the opposite dendrogramme branch. The same effect, however, is not as pronounced when examining the results from the *Tabula Muris* dataset (Figure 4.3C), where cell type sampling is less biased per tissue. Tissues with similar levels of expression of the same markers can be observed to cluster together (heatmap tissue clusters with AU p-value<=95: Spleen and Marrow; Trachea to Muscle; Kidney and Liver; Brain_Neurons and Pancreas); however the stark immune/non-immune division observed for human data is no longer present. We can thus conclude that tissue similarity, as defined by cell type correspondence, is driven by cell identity, in particular by the major lineage (immune, epithelial, endothelial), yet can be affected by cell type sampling proportions.

Tissue identity is also reflected in gene expression, and therefore in the genes with the top coefficients determined by the *CellTypist* model. To unbiasedly probe the existence of tissue-specific signatures in the top genes of all clusters, tissue signatures were derived from bulk RNA-seq data, using data from the GTEx Consortium for human (Consortium, 2015) and from the ENCODE Consortium for mouse (Dunham et al., 2012) (see Methods Section 4.4.2). This provided independent references for tissue identity using gene expression. Inspection of human tissue identity enrichment in cell clusters per tissue (Figure 4.3B) shows that, despite the sets of tissues in the *CellTypist* and GTEx datasets not overlapping completely, matching between them is mostly concordant. Most immune cell-enriched tissues cluster independently (AU p-value = 96 for both branches) by having many clusters enriched in blood and spleen-specific genes. Beyond this separation, There is a high correspondence between tissue-specific genes and the tissues present in the data. Examples are Liver, Brain, Testis, Lung (Parenchyma), Kidney, Pancreas, and Colon (matching Small Intestine). Among the tissues with more diverging matching are Skin, likely because of the very biased cell sampling (Treg and Tmem cells (Miragaia et al., 2019)). Other tissue correspondences might derive from functional similarities, such as Pancreas

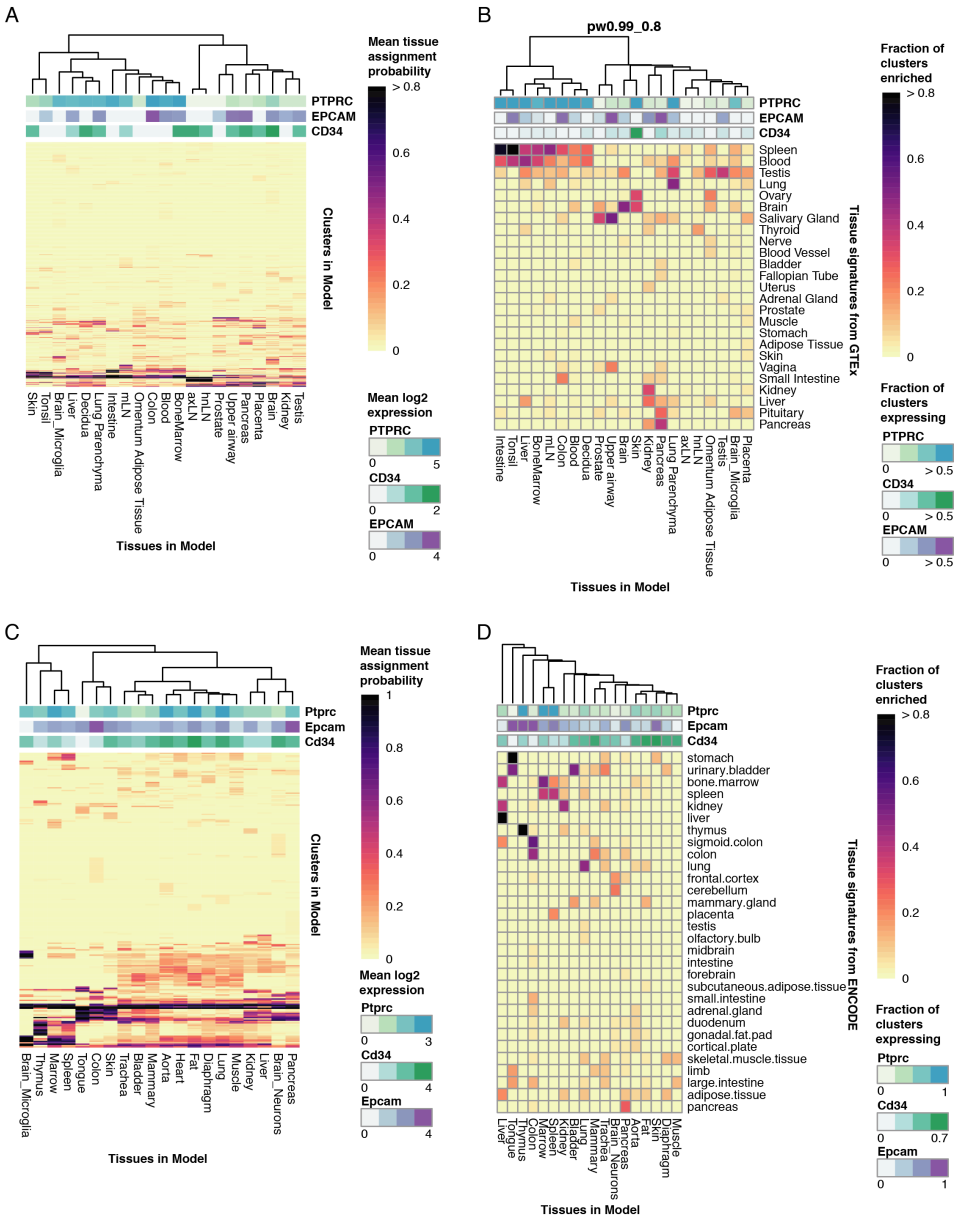


Fig. 4.3: **Cell identity relationships across tissues**
(A) Heatmap of the mean assignment probability of cells from a per-tissue cluster to the clusters of each given tissue in the human collection dataset. **(B)** Heatmap of the fraction of human collection clusters from a given tissue whose *CellTypist* model (thr1 = 0.99; thr2 = 0.8, see Chapter 3 Section 3.3.2) signature is enriched in certain tissue-specific genes. Gene signatures were derived from GTex bulk RNA-seq data. **(C)** Heatmap of the mean assignment probability of cells from a per-tissue cluster to the clusters of each given tissue in the Tabula Muris dataset. **(D)** Heatmap of the fraction of Tabula Muris clusters from a given tissue whose *CellTypist* model (thr1 = 0.8; thr2 = 0.99, see Chapter 3 Section 3.3.1) signature is enriched in certain tissue-specific genes. Gene signatures were derived from ENCODE tissue bulk RNA-seq data.

and Pituitary (hormonal secretion), Tonsil and Spleen (lymphoid tissues), or Upper Airway and Vagina (mucosal epithelia).

Similar specificity relationships can be observed in the *Tabula Muris* dataset (Figure 4.3D), with a high matching for Pancreas, Liver, Kidney, Bladder, Thymus, and Lung, among others. Matching by functional or cell composition similarity was also present, such as Fat and mammary gland, or Diaphragm and skeletal muscle tissue. However, the evident division between immune/non-immune visible in human was once again absent. This further indicates that comparative analysis of tissue composition at the single cell level must be based on datasets that are representative of the tissues' cellular composition, in order to avoid biased characterisations.

The tissue relationships highlighted by the gene set enrichment directly derive from the *CellTypist* model trained and the cell groupings that the pipeline defines. Examining other model alternatives shows that in some of them the tissue hierarchy is maintained (Figure C.3), with the exception of the thr1 = 0.1, thr2 = 0.1 model. This is likely caused by excessive merging of clusters, leading to non-meaningful groupings and not so meaningful gene coefficients from the model.

Plotting the clusters resulting from cross-tissue merging in *CellTypist* can also reveal the similarity across tissues (Figure C.4). As already shown by Figure B.3A, the model with thr1 = 0.99 and thr2 = 0.8 is the one with the lowest number of merged clusters. We can however still observe clusters merging across tissues that have similar profiles and were included in the "immune enriched" group in Figure 4.3B - liver and bone marrow, lung parenchyma and intestine, decidua and omentum adipose tissue - as well as tissues that have functional associations - decidua and placenta, upper airway and lung parenchyma. The remaining models appear to maintain the occurrence of these associations between tissues, like the close clustering between axLN and hnLN, or the association of tissues including more immune sampling with blood and bone marrow. This is further underscored when the tissue gene signatures are examined in the merged clusters of each model (Figure C.5). Both thr1 = 0.4 and thr2 = 0.99 and thr1 = 0.25 and thr2 = 0.25 again present the distinctive pattern of clustering the tissue signatures by the tissue functions as described before for the immune/non-immune partitions. The first model (thr1 = 0.99 and thr2 = 0.8) also shows some of this pattern, although not as evidently, likely due to the lower number of merged clusters. The same can not be observed for the thr1 = 0.1 and thr2 = 0.1 model, likely due to excessive merging leading to a less meaningful classifier.

Together, these results show that single cell populations in different tissues capture some of the tissue biology and specificity, representing functional and compositional relationships between them.

4.2.3 Gene expression hallmarks of cell identity

The training of a logistic regression-base classifier model as used in *CellTypist* allows for a direct evaluation of the genes important for the classification of each cluster through their learned coefficients. With a comprehensive cell type reference, we can start to unravel what are the key determinants of cell identity across tissues.

Relationships between human cell clusters were probed by counting the number of pairwise shared genes. The top 500 genes were used to avoid a hard coefficient value threshold across clusters, since the top values can be very variable between them. Clustering once more revealed a division between most immune and non-immune clusters (Figure 4.4A). Moreover, various clusters containing cells from the same tissue were also grouped together, hinting at the existence of gene expression programmes shared by the different cell types within a tissue.

The concept of "cell type" is defined in different ways by different biomedical research communities, yet it is consistently related to a cell's molecular phenotype, i.e. the molecules involved in cellular function. These can either be the effector molecules directly responsible for the cell's array of functions, or the genomic regulators controlling the expression of genes involved in these functions. It has been showed that tissue-specificity at the gene expression level is mostly due to transcription factor-gene regulatory interactions (Sonawane et al., 2017). The *CellTypist* model was used to assess what types of genes were more often at the top of the model coefficient rankings, which reflect the importance of expression of that gene in classifying a cell type. The following gene categories were considered (see Methods Section 4.4.2): Transcription Factors, Chromatin Modulators, Kinases and Phosphatases, Ligases and Deubiquitinases, Catalytic enzymes, Housekeeping genes, Receptors, Secreted proteins, Transmembrane proteins, and Peripheral membrane proteins.

Both in human (Figure C.6) and mouse (Figure C.7), we did not observe a large difference between the mean expression levels of genes from different groups, or between their maximum coefficients; most highly ranked genes (in the top 500) had a mean expression level around 10 reads. This coincided with a high correlation (0.56 in human, 0.86 in mouse, Spearman correlation coefficient) between mean expression and maximum reported coefficient, suggesting a dependency of gene

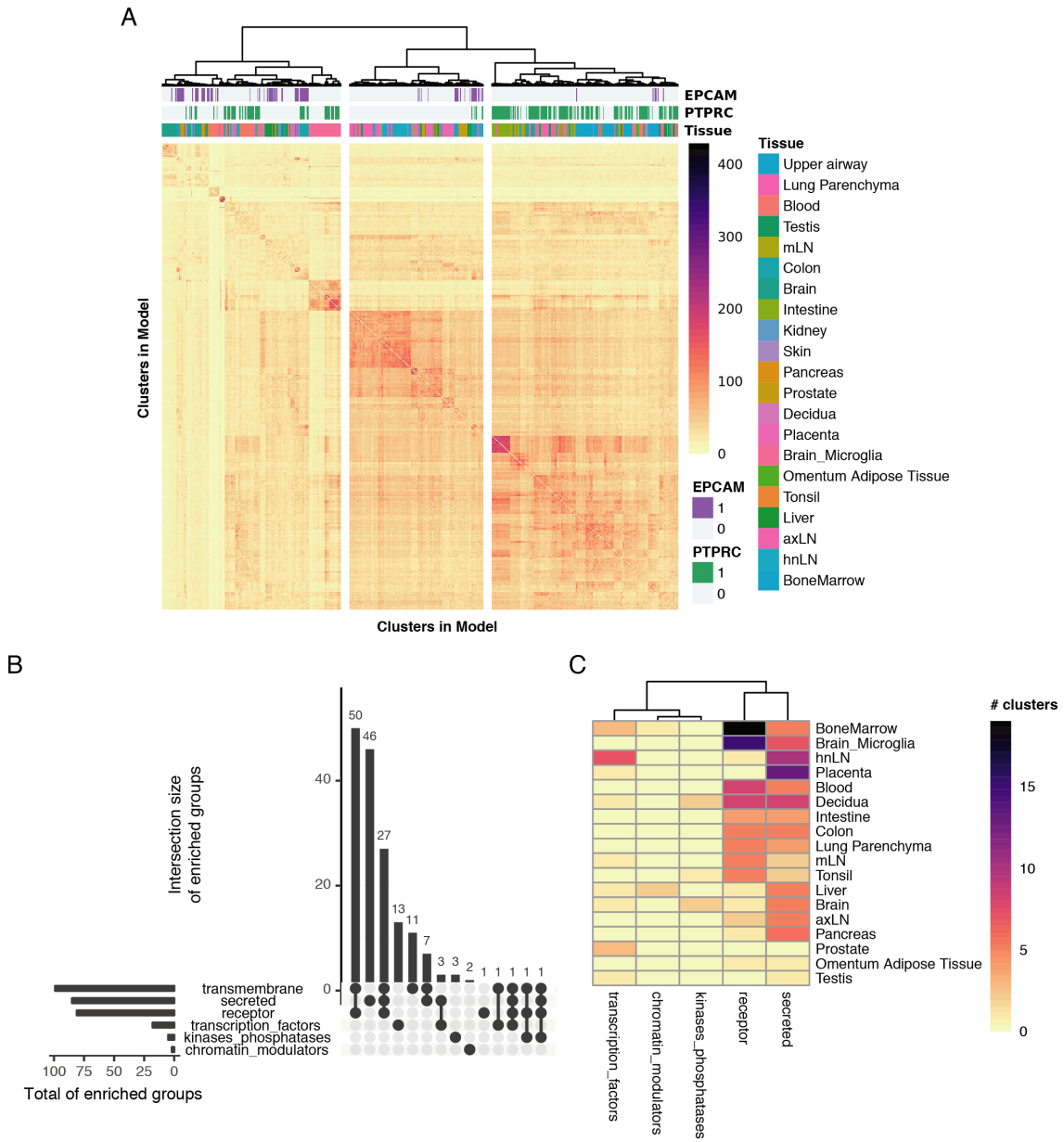


Fig. 4.4: **Top gene groups for cell identification across human tissues**
(A) Clustered heatmap of the number of genes in common between pairs of *CellTypist* clusters (thr1 = 0.99, thr2 = 0.8) in human data. Genes per cluster were determined as those with the top 500 coefficients learned by the model. Values in the diagonal (number of genes per cluster, 500) were set to 0. **(B)** Upset plot counting the number of clusters enriched for a group of genes with a specific function. **(C)** Heatmap of number of clusters per tissue (y-axis) enriched for groups of genes with a specific function (x-axis). For panels (B) and (C), the gene groups "transcription factors", "transmembrane", "secreted", "receptors", "membrane peripheral proteins", "kinases and phosphatases", "chromatin modulators", "catalytic enzymes", "housekeeping genes" were tested. Only the terms enriched in at least one cluster are shown.

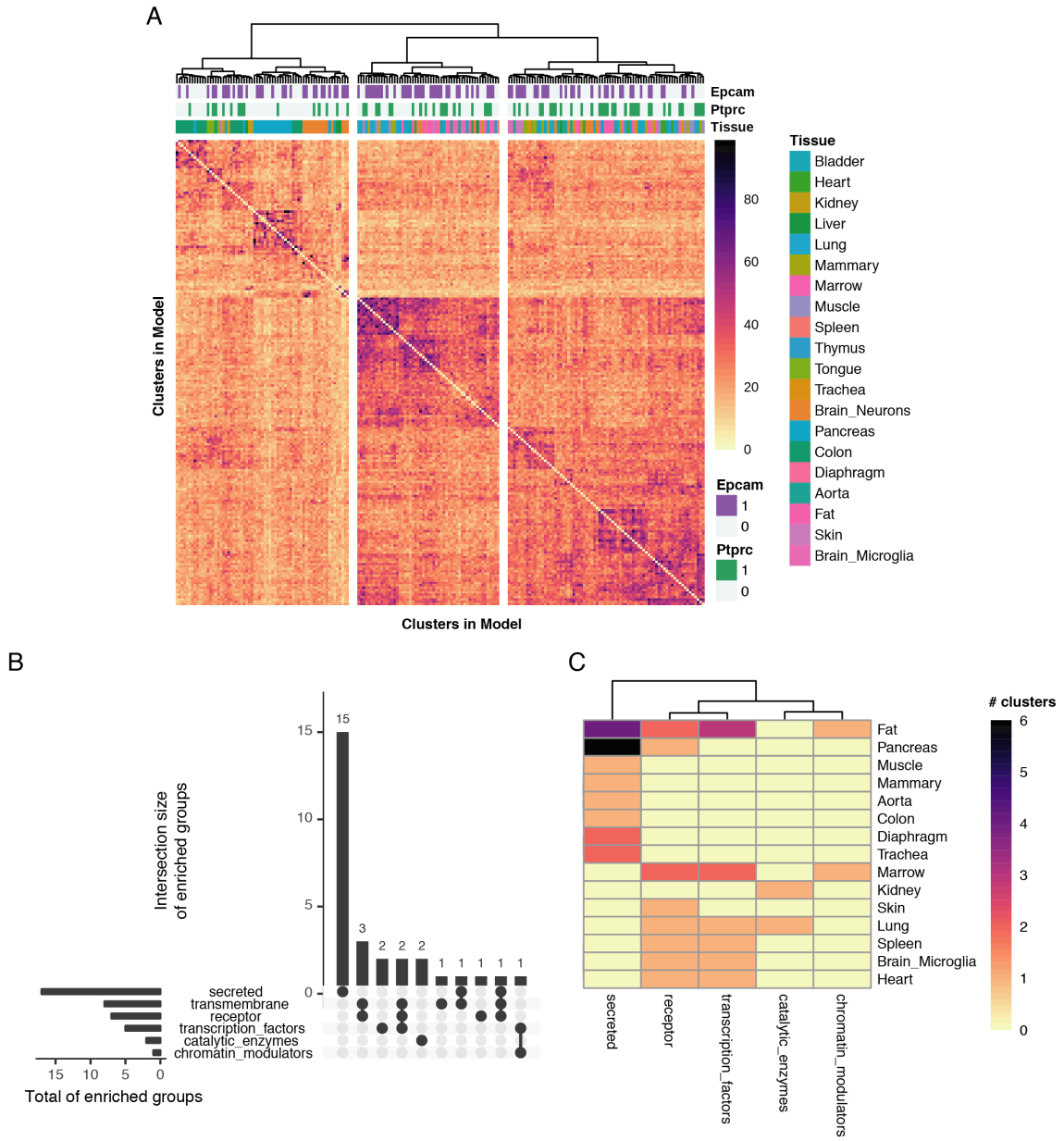


Fig. 4.5: **Top gene groups for cell identification across mouse tissues**
(A) Clustered heatmap of the shared number of genes between pairs of *CellTypist* clusters (thr1 = 0.8, thr2 = 0.99) in the *Tabula Muris* data. Genes per cluster were determined as those with the top 500 coefficients learned by the model. Values in the diagonal (number of genes per cluster, 500) were set to 0. **(B)** Upset plot counting the number of clusters enriched for a group of genes with a specific function. **(C)** Heatmap of number of clusters per tissue (y-axis) enriched for groups of genes with a specific function (x-axis). For panels (B) and (C), the gene groups "transcription factors", "transmembrane", "secreted", "receptors", "membrane peripheral proteins", "kinases and phosphatases", "chromatin modulators", "catalytic enzymes", "housekeeping genes" were tested. Only the terms enriched in at least one cluster are shown.

importance for classification on expression level. However, this relationship appears to be non-linear, as evidenced by its shape, which remains constant for genes with about 10 reads or more, and by the low Pearson Correlation Coefficient (0.05 in human, 0.08 in mouse). Thus, gene expression level only affects the learned model coefficient when comparing lowly and highly expressed genes.

Testing the gene groups described above for enrichment (see Methods Section 4.4.4) showed a consistent pattern for all surveyed models of predominantly enriched membrane and secreted proteins (Figure 4.4B, Figure C.8). A number of clusters also had transcription factors enriched in their top hits, albeit in markedly lower number. Enrichment for the tested gene groups appeared evenly distributed across tissues, and did not group them in any meaningful manner (Figure 4.4C). Lastly, it is also notable that only a fraction of the total clusters showed enrichment for any of the classes tested, which could be due to the restrictive test that only looks for enrichment at the very top genes, as well as the non-comprehensive list of functions tested.

Examining the model produced by *CellTypist* on the *Tabula Muris* dataset revealed similar results. The grouping of immune versus non-immune clusters present in human was again absent (Figure 4.5A), as had been observed in the previous Section (Figure 4.3). The patterns for gene groups were nonetheless similar, with a greater enrichment of secreted proteins across all cell clusters (Figure 4.5B), and the largest significant groups spread across more various tissues (Figure 4.5C).

These results point to the greater importance of the gene expression regulatory network's output molecules (genes coding for membrane and secreted proteins), when computationally defining the identity of a cell.

4.3 Discussion

From its inception, the Human Cell Atlas (HCA) consortium has aimed to "define all human cell types in terms of distinctive molecular profiles (such as gene expression profiles)" (Regev et al., 2017), a task that can not be easily accomplished by a single team. Beyond the financial and ethical constraints, collecting good quality scRNA-seq data requires tissue-specific knowledge, as well as profiling using both top-down and bottom-up approaches to obtain an overview of cell populations, while capturing cell type-specific phenotypic variations. Yet as data on human cells accumulates, methods capable of compiling the cellular census envisioned by the HCA members, and making it available to the community will be of great use.

The human data presented provides a broad overview of several organs. This leads the cell type reference generated by *CellTypist* to be broadly applicable to new datasets. This reference is dependent on the way these tissues are sampled. Currently, many of them are mostly or totally composed of immune cells which, while adding valuable information about their diverse phenotypes, can also bias the model. Collecting more datasets is the ideal way of mitigating this problem, but it can also be addressed by using data augmentation or downsampling approaches (Hie et al., 2019b; Wong et al., 2016b). This would be especially relevant at the model training step, as we have observed the clear impact of number of cells per label in classification accuracy (Figure 3.4C, Figure 3.5C, Figure 3.7D).

Consistent data integration is also essential to avoid redundant classes and misleading interpretations about cell type and tissue relationships. Data integration for scRNA-seq is still a heavily studied topic (Haghverdi et al., 2018; Lopez et al., 2018; Polański et al., 2019; Stuart et al., 2019), and can considerably influence the cell groupings detected in the data. *CellTypist* is likely to evolve as a pipeline, in order to adopt a within- and cross-tissue integration framework that closely reflects the cell type information available for each dataset. This integration will also lead to a clear cell type label for the model, while also reflecting the cell type resolution limitations of the classifier.

Tissue identity relationships appear as an emergent result from the application of *CellTypist*. The associations revealed between tissues are present at the cross-tissue integration stage (Figure 4.3A), and then also reflected in the top genes learned by the logistic regression model (Figure 4.3B). Furthermore, tissue identity is to some degree robust to incorrect or excessive grouping of single-cells (Figure C.3), which reveals that tissues-specific expression programmes might be intrinsic to the core cell identity. The resolution of these tissue connections and programmes can be improved by broader cell type sampling and integration. This will allow the model to reveal a more fine-grained hierarchy beyond the immune/non-immune split, and ultimately map cellular phenotypes to a structured cell identity atlas.

The data compiled offers for the first time a window into the gene expression hallmarks of cell identity for the first time. Analysis of enriched gene expression programmes can be improved by using a more uniform gene reference, as well as adopting more informative labels for the clusters obtained (which can come from improved dataset merging or manual annotation). Nonetheless, the analysis showed consistently ranked receptors and secreted molecules above transcription factors when defining cell identity (Figure 4.4B, Figure C.8). This is in agreement with

previous reports (Sonawane et al., 2017), yet this is the first instance where this type of analysis could achieve this level of cell type resolution. Importantly, defining which genes make up the core of cellular phenotypes is not the same as defining cell identity regulation. However, knowledge of the minimal gene expression set required to classify or obtain a determined phenotype (and consequently function) is a key point in understanding the operational definition of cell types. Thus, the expansion and improvement of the *CellTypist* reference will increasingly provide a foundation to understanding how cell types arise and evolve (Zimmermann et al., 2019), and will help prioritise gene targets for effective cellular engineering.

This large human cell type reference can be very useful to characterise cell identity in a variety of systems. In disease-focused studies, the steady-state reference provided by *CellTypist* can automatically annotate the cells obtained from a disease sample, without relying on a matching healthy sample. This is useful in large scale studies that aim to quantify cell number alterations in disease, yet steady-state cells would still be required to identify disease-specific gene expression programmes or cell subpopulations. Another potential use is to characterise cell fates and heterogeneity when differentiating organoids. Classifying scRNA-seq data from the generated organoids using an unbiased reference can reveal the cell types present that a specific protocol was able to differentiate. *CellTypist* will also be available as an online resource, where the model can be directly used, and is accompanied by a database showing the defining characteristics of each cell type - marker genes detected, tissues of origin, datasets characterising them, and similar cell types. This is further intended to be articulated with a Cell Ontology (Bard et al., 2005), and have cell names be consistently used when new data is produced, with a direct correspondence to both databases. Lastly, future releases of *CellTypist* models will include more species, adding an evolutionary layer to our knowledge of cell identity.

4.4 Methods

4.4.1 *CellTypist* parameter optimisation and training

Use of the integration and model training pipeline in the human dataset collection was done as described in Chapter 3 Section 3.3.2, and is again briefly explained here. Data from the same tissues was integrated and clustered using the Leiden algorithm (Traag et al., 2019) at several resolutions. For tissues with cell type annotations, resolution was optimised using the split-join distance (Dongen, 2000)

between clusters and cell type annotation and constrained to a number of clusters at least as large as the number of cell type annotations in the largest collected dataset (Figure 3.7A).

Following clustering, per tissue logistic regression models were trained, running for 10 epochs of a maximum of 100 iterations each. These models were used to run the cross-tissue cluster merging pipeline (Chapter 3 Section 3.2.2), and a combination of parameters was chosen based on the ratio of split-join distances (merged vs annotated cell types over per tissue vs annotated cell types) (Figure 3.7B), resulting in the choice of thr1 = 0.99 and thr2 = 0.8. Additionally, three other combinations were chosen for comparison: thr1 = 0.4 and thr2 = 0.99, the combination with the top split-join ratio when only considering merged clusters (Figure B.3C, Figure B.5A-B); thr1 = 0.25 and thr2 = 0.25, one of the combinations with the highest fraction of merged clusters (Figure B.3B, Figure B.5C-D); thr1 = 0.1 and thr2 = 0.1, the combination with the highest fraction of merged clusters, as well as highest split-join fraction (Figure B.3B, Figure B.5E-F).

The groupings obtained were used to train a logistic regression model using Stochastic Gradient Descent (Chapter 3 Section 3.2.3). Training was done for 25 epochs of a maximum of 100 iterations each, where in each iteration 1000 cells were seen by the model. 90% of the total data was used as a training set, and the remaining as a left out test set that was tested at every iteration (Figure 3.7C-D, Figure B.5).

4.4.2 Obtaining gene group lists

The groups of genes here presented were chosen to reflect various broad functions present in cells. They are not exhaustive, and overlaps between gene sets exist due to the ambiguity of some categories. In some tests, various categories were used, yet only those with at least one positive result were reported (Figure 4.3B, Figure 4.4B).

Cell type markers (from (Madissoon et al., 2019)): for each tissue, the function "rank_genes_groups" from scanpy (Wolf et al., 2018) was used to determine the markers of each cell type. A filter of q-value<=0.01 and log2 fold-change>=1 was used to select the top markers of each annotated group.

GO Terms: GO Terms were downloaded using the biomaRt R package (Durinck et al., 2009). Genes from different terms were then grouped in the following categories (similar to (Hagai et al., 2018)): chromatin modulators (GO:0006338 (chromatin remodelling), GO:0003682 (chromatin binding), GO:0042393 (histone

binding), and GO:0016568 (chromatin modification)); kinases and phosphatases (GO:0004672 (protein kinase activity) and GO:0004721 (phosphoprotein phosphatase activity)) and catalytic enzymes (GO:0003824 (catalytic activity)).

Transcription Factors: Human transcription factors were obtained from AnimalTFDB v3.0 (http://bioinfo.life.hust.edu.cn/AnimalTFDB/) (Hu et al., 2019).

Housekeeping genes: Housekeeping genes were obtained from https://m.tau.ac.il/~elieis/HKG/ (Eisenberg and Levanon, 2013).

Cell communication-associated genes: Genes involved in cell-cell communication were obtained from cellphonedb.org (Efremova et al., 2019). Only genes annotated as "transmembrane", "secreted", "peripheral", and "receptor" were kept. Given the structure of the annotation in this database, some genes are included in more than one group. In particular, most receptors and some secreted proteins are also classified as transmembrane.

Tissue-specific genes: Tissue specific genes were determined as described in (Yanai et al., 2005) (see (Kryuchkova-Mostacci and Robinson-Rechavi, 2017) for a benchmark). Briefly, RNA-seq expression data from the GTex Consortium (human, https://gtexportal.org/home/index.html) or ENCODE Consortium (mouse, https://www.encodeproject.org/) were obtained (Consortium, 2015; Dunham et al., 2012). The tau statistic was calculated for each gene, and it consists on the normalised deviation of a gene's expression in a tissue from the maximum expression value observed. Only genes with a tau value greater than or equal to 0.5 were kept C.1. This threshold was used in order to have enough genes per group to test tissue specificity. Despite this being a very relaxed threshold, no genes shared between tissues were found. Moreover, using a more restrictive threshold like 0.9 resulted in numbers within the same order of magnitude of genes for each tissue, although not enough to test for enrichment.

4.4.3 Clustering

Clustering (in heatmaps) was performed using the hclust function from R, with euclidean distance and the "ward.D2" method. Clustering uncertainty was assessed using the pvclust R package, and AU p-values greater than or equal to 95 were considered significant.

4.4.4 Enrichment of gene groups

To obtain enriched groups of genes (Sections 4.2.2 and 4.2.3), the top 500 genes based on their model coefficients were obtained for each cluster. Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was performed using the liger R package (https://cran.rstudio.com/web/packages/liger/index.html), considering the gene sets as defined in Section 4.4.2. Enrichment was deemed signifficant if the q-value was lower than 0.05, and if the enrichment score was positive, signifying an enrichment in the top genes. In heatmaps plotting GSEA results (Figure 4.3B-D; Figure C.3), the colour scale is capped at 0.8 (fraction of enriched clusters per tissue), and the annotation scales are capped at 0.5 (fraction of clusters with mean expression of the indicated gene of at least 1). Clusters merge across tissues were only counted towards the tissue contributing the most cells to them.

Chapter 5

Concluding remarks

Developments in single-cell genomics are still shaping the way we define cellular identity. With the increasing number of cell types, organs and species profiled, we are bound to obtain an exhaustive overview of eukaryotic cell diversity, together with their genomic determinants. This work illustrates the importance of studying cell types across different tissues, and discussed computational challenges as well as solutions for the integrative atlasing of cellular diversity.

5.1 Cells and genes trade-offs in single-cell profiling

The number of cells profiled per study is still increasing exponentially (Svensson et al., 2018). This has been accompanied by a marked expansion in the number of studies using single-cell technologies (Svensson and Beltrame, 2019), much of it due to the spread in use of a more standardised cell isolation and sequencing pipeline, 10x Genomics' Chromium technology. This democratisation of single-cell omics is resulting in more cell types, tissues and species being profiled. Nevertheless, single-cell studies should be designed with a clear goal, and the choice of protocol should be adequate to the question at hand.

With regards to the type of sequencing, scRNA-seq protocols can be broadly split between full length transcript profiling and 5'/3' RNA tagging. Full length protocols - the most widely used being Smart-seq2 (Picelli et al., 2014) - follow in the footsteps of the majority of bulk RNA-seq studies. Smart-seq2 is still dependent on mRNA isolation by the poly-A tail, and thus does not reveal changes in non-polyadenylated transcript as other protocols might (Hayashi et al., 2018; Verboom et al., 2019). Despite this, Smart-seq2's full length characteristics have been important to study immune cells.

The development of TraCeR (Stubbington et al., 2016) and BraCeR (Lindeman et al., 2018) have allowed the detection of TCR and BCR transcripts in single-cell data, which in turn have been used to lineage trace T cells with similar developmental origin (Lönnberg et al., 2017) and Treg cell migration between tissues (Miragaia et al., 2019) (Chapter 2). Smart-seq2 has also allowed uncovering the diversity of KIR receptors in NK cells at the maternal-fetal interface (Vento-Tormo et al., 2018). Splicing-oriented studies with this protocol, on the other hand, have been scarce (Arzalluz-Luque and Conesa, 2018), yet splicing can be important in revealing important features of cell identity. Combination of Chromium and PacBio long read sequence has revealed cell type specific isoforms in mouse cerebellum (Gupta et al., 2018). Other changes in isoform usage also exist that can influence cell identity, yet this has been underappreciated.

The dominance of 3' and 5' sequencing protocols stems from the fact that a large number of cells is more important to revealing cell diversity in a given tissue or condition than increased sequencing depth or number of genes per cell (Svensson et al., 2019), which has been showed early on when profiling bipolar retinal cells in mouse (Shekhar et al., 2016). Droplet-based protocols allow the user to more easily isolate a large number of cells, which are then only sequenced at lower levels. This increase in cell numbers was necessary in the Treg cell work presented in Chapter 2 to detect the subpopulations composing the lymph node-peripheral tissue trajectory (Figures 2.2 and 2.3). Thus, at the transcriptomic level, different protocols can serve complementary functions - either increasing the resolution of the cellular census, or providing a more detailed representation of the molecular makeup of cell populations.

5.2 Building a transcriptomic atlas of cell types

The study presented in Chapter 2 shows that, to unravel the full extent of cell identity, it is not enough to unbiasedly profile a tissue, since even low-frequency cell populations may reveal functional heterogeneity. Furthermore, the relevance of this is sometimes only apparent once more tissue-specific context is added. In isolation, the census of colonic Treg cells would only reveal different levels of activation, but once this was combined with the draining mesenteric lymph node (mLN) populations, and compared with Treg cells in the brachial lymph nodes, it became clear that these subpopulations formed a continuum across organs, and the subpopulations present in the mLN expressed genes that coded for homing chemokine receptors specific for the colon.

The development of single-cell sequencing methods has unlocked the ability to perform unbiased cellular phenotyping. Yet there are several layers, from DNA, RNA and protein, to probe this phenotype. From these, at the single-cell level, RNA is by far the most widely available. While it is not as close to cellular function as proteins, it is a good approximation, and can be unbiasedly amplified. The spread of cellular transcriptomic profiling was not initially accompanied by a development of dedicated databases for this type of data, although more recent efforts have been made towards this end (Alavi et al., 2018; Franzén et al., 2019), and it is the goal of the Human Cell Atlas to gather and standardise single-cell expression data. Moreover, most of the data produced is not accompanied by cell type annotations in a machine-readable format, nor does it follow a standardised nomenclature. This is likely because the existing ontologies (Bard et al., 2005) were not prepared for this explosion in cell profiling and the diversity of cell types and states it brought. Thus, the existence of these scRNA-seq datasets creates an opportunity to develop and update an informed cell type reference (Aevermann et al., 2018). Chapter 3 introduced *CellTypist*, a method to integrate scRNA-seq data from multiple sources and tissues. This pipeline does not require a uniform annotation *a priori*, and produces an interpretable model for annotation of new data.

While most cell population profiling focuses on RNA, other aspects are also relevant. Open chromatin regions, which can be identified through scATAC-seq, are often involved in regulation of gene expression, and have been shown to be sufficient to distinguish cell types similarly to expression profiling (Cusanovich et al., 2018). An open chromatin cell type atlas can then provide a more regulatory perspective on cell identity, perhaps more clearly illustrating what effective alterations at the DNA level result in acquisition or loss of cellular phenotypes.

Cell type references like *CellTypist* have a multitude of applications, ranging from basic science to applied biomedicine. The predictive capabilities of this sort of models can be used to test cellular responses in organoids (Brazovskaja et al., 2019). This assessment can range from evaluating the differentiation potential of cultured cells, to measuring deviations and responses caused by external factors, like varying differentiation molecules or infectious agents. Ultimately, this can further improve the efforts in the field of tissue engineering, by guiding the development of *in vitro* differentiation protocols (Camp et al., 2018). Likewise, these references can also be used in a clinical setting to probe changes in cell diversity in disease. In cancer, infiltration and phenotypic changes of immune cells can be assessed, and single-cell phenotyping of tumour cells can monitor its progression. Monitoring of

cell abundances and diversity in the clinic can provide a new view of disease from a "cell ecology" perspective.

More fundamentally, *CellTypist*, as an integrated, cross-tissue cell type compendium can inform us on the key genes that define the core cellular phenotype.

5.3 Defining cellular identity

The advent of single-cell genomics has re-ignited the debate on the definition of cell type identity. Historically, cell types have been defined based on their morphology, location, function, or developmental origin. Development of cellular staining, and especially flow cytometry, have added molecular phenotyping to this list. While flow cytometry already offered a large cell throughput capable of detecting even the smallest populations, the revolutionary aspect of single-cell transcriptomics has been the unbiased probing of RNA molecules, revealing a new high-resolution cellular map of gene expression programmes.

It is only through integration that we can achieve a organism-scale picture of cell identity. This is achieved by *CellTypist*, which is capable of resolving cell type correspondences across tissues (Figure 3.1, Figure 4.3), and provides the list of genes at the core of each cell grouping (Figure 4.4). Despite the discussed limitations, owed in part to the still limited diversity of data available, *CellTypist* lays the groundwork and reveals the first systematic picture of human cell types (with an expansion to other species in sight).

The transcriptomic composition of cells is vastly informative for their taxonomy, yet only makes up a small portion of the information we can obtain. Other omics modalities (open chromatin, chromatin modifications, methylation, proteomics, ...) can provide equally informative yet complementary perspectives on cellular identity. Furthermore, these can be integrated computationally (Stuart et al., 2019) or obtained simultaneously using appropriate protocols (Angermueller et al., 2016; Clark et al., 2018). Nonetheless, an ideal compendium of cell types should strive to go beyond this low level and invasive characterisation, and merge back into the knowledge obtained from other modalities. Cellular interactions are of great importance to cell function, and thus spatial information adds a relevant layer to this. Mapping the developmental trajectories of all cell types can inform us on their origin and generative processes. Morphology is the most easily observed characteristic, and heavily related to cell function, thus controlled by the genome. Only through integration can this systems view of cell biology come to fruition.

When possessing information on these many layers of cell phenotypes, we will be able to more accurately define the boundary between cell types and cell states. Often these can be observed in each individual modality - transient versus definitive cell shapes, immune lineages and their response to pathogens, or intermediate versus leaf stages in cellular differentiation. Yet these perspectives need each other, as cellular form and function should be understood in the context of its origin and genomic programming. Reconciling these different perspectives through a multi-window approach will provide us with a complete blueprint of the basic unit of life. It is expected that the unified view provided by the Human Cell Atlas - and indeed all cell atlases - results in a Modern Synthesis of Cell Theory.

References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, 20(1):194.

Adam, M., Potter, A. S., and Potter, S. S. (2017). Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. *Development*, 144(19):3625–3632.

Aevermann, B. D., Novotny, M., Bakken, T., Miller, J. A., Diehl, A. D., Osumi-Sutherland, D., Lasken, R. S., Lein, E. S., and Scheuermann, R. H. (2018). Cell type discovery using single-cell transcriptomics: implications for ontological representation. *Human Molecular Genetics*, 27(R1):R40–R47.

Agace, W. W. (2006). Tissue-tropic effector T cells: generation and targeting opportunities. *Nature Reviews Immunology*, 6(9):682.

Alavi, A., Ruffalo, M., Parvangada, A., Huang, Z., and Bar-Joseph, Z. (2018). A web server for comparative analysis of single-cell RNA-seq data. *Nature Communications*, 9(1):4768.

Ali, N., Zirak, B., Rodriguez, R. S., Pauli, M. L., Truong, H.-A., Lai, K., Ahn, R., Corbin, K., Lowe, M. M., Scharschmidt, T. C., Taravati, K., Tan, M. R., Ricardo-Gonzalez, R. R., Nosbaum, A., Bertolini, M., Liao, W., Nestle, F. O., Paus, R., Cotsarelis, G., Abbas, A. K., and Rosenblum, M. D. (2017). Regulatory T Cells in Skin Facilitate Epithelial Stem Cell Differentiation. *Cell*, 169(6):1119–1129.e11.

Alquicira-Hernández, J., Sathe, A., Ji, H. P., Nguyen, Q., and Powell, J. E. (2018). scPred: Cell type prediction at single-cell resolution. *bioRxiv*, page 369538.

Altin, J. G. and Sloan, E. K. (1997). The role of CD45 and CD45-associated molecules in T cell activation. *Immunology and Cell Biology*, 75(5):430–445.

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., and Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13(3):229–232.

Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., Goga, A., Sirota, M., and Butte, A. J. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature Communications*, 8(1):1077.

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., and Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163.

Arzalluz-Luque, A. and Conesa, A. (2018). Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biology*, 19(1):110.

Bagnoli, J. W., Ziegenhain, C., Janjic, A., Wange, L. E., Vieth, B., Parekh, S., Geuder, J., Hellmann, I., and Enard, W. (2018). Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nature Communications*, 9(1):1–8.

Bandura, D. R., Baranov, V. I., Ornatsky, O. I., Antonov, A., Kinach, R., Lou, X., Pavlov, S., Vorobiev, S., Dick, J. E., and Tanner, S. D. (2009). Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. *Analytical Chemistry*, 81(16):6813–6822.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Çolak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J., and Blencowe, B. J. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*, 338(6114):1587–1593.

Bard, J., Rhee, S. Y., and Ashburner, M. (2005). An ontology for cell types. *Genome Biology*, 6(2):R21.

Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K., and Kharchenko, P. V. (2019). Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nature Methods*, 16(8):695–698.

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., and Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, 3(4):346–360.e4.

Barreiro, L. B. and Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics*, 11(1):17–30.

Behjati, S., Lindsay, S., Teichmann, S. A., and Haniffa, M. (2018). Mapping human development at single-cell resolution. *Development*, 145(3):dev152561.

Bernstein, M. N. and Dewey, C. N. (2019). Hierarchical cell type classification using mass, heterogeneous RNA-seq data from human primary cells. *bioRxiv*, page 634097.

Bettelli, E., Carrier, Y., Gao, W., Korn, T., Strom, T. B., Oukka, M., Weiner, H. L., and Kuchroo, V. K. (2006). Reciprocal developmental pathways for the generation of pathogenic effector TH17 and regulatory T cells. *Nature*, 441(7090):235–238.

Bjorklund, A. K., Forkel, M., Picelli, S., Konya, V., Theorell, J., Friberg, D., Sandberg, R., and Mjösberg, J. (2016). The heterogeneity of human CD127$^+$ innate lymphoid cells revealed by single-cell RNA sequencing. *Nature Immunology*, 17(4):451–460.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bollrath, J. and Powrie, F. M. (2013). Controlling the frontier: regulatory t-cells and intestinal homeostasis. *Semin. Immunol.*, 25(5):352–357.

Bonner, W. A., Hulett, H. R., Sweet, R. G., and Herzenberg, L. A. (1972). Fluorescence Activated Cell Sorting. *Review of Scientific Instruments*, 43(3):404–409.

Boufea, K., Seth, S., and Batada, N. N. (2019). scID: Identification of equivalent transcriptional cell populations across single cell RNA-seq data using discriminant analysis. *bioRxiv*, page 470203.

Braga, F. A. V., Kar, G., Berg, M., Carpaij, O. A., Polanski, K., Simon, L. M., Brouwer, S., Gomes, T., Hesse, L., Jiang, J., Fasouli, E. S., Efremova, M., Vento-Tormo, R., Talavera-López, C., Jonker, M. R., Affleck, K., Palit, S., Strzelecka, P. M., Firth, H. V., Mahbubani, K. T., Cvejic, A., Meyer, K. B., Saeb-Parsy, K., Luinge, M., Brandsma, C.-A., Timens, W., Angelidis, I., Strunz, M., Koppelman, G. H., Oosterhout, A. J. v., Schiller, H. B., Theis, F. J., Berge, M. v. d., Nawijn, M. C., and Teichmann, S. A. (2019). A cellular census of human lungs identifies novel cell states in health and in asthma. *Nature Medicine*, 25(7):1153–1163.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348.

Brazovskaja, A., Treutlein, B., and Camp, J. G. (2019). High-throughput single-cell transcriptomics on organoids. *Current Opinion in Biotechnology*, 55:167–171.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095.

Brown, R. (1866). On the Organs and Mode of Fecundation of Orchidex and Asclepiadea. In *Miscellaneous Botanical Works, Vol. I*. London: Ray Society.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. (2017). f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology*, 18(1):212.

Burzyn, D., Kuswanto, W., Kolodin, D., Shadrach, J. L., Cerletti, M., Jang, Y., Sefik, E., Tan, T. G., Wagers, A. J., Benoist, C., and Mathis, D. (2013). A Special Population of Regulatory T Cells Potentiates Muscle Repair. *Cell*, 155(6):1282–1295.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420.

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, 16(1):43–49.

Camp, J. G., Wollny, D., and Treutlein, B. (2018). Single-cell genomics to guide human stem cell and tissue engineering. *Nature Methods*, 15(9):661–667.

Campbell, D. J. and Koch, M. A. (2011). Phenotypical and functional specialization of FOXP3+ regulatory T cells. *Nat. Rev. Immunol.*, 11(2):119–130.

Campbell, K. R. and Yau, C. (2017). switchde: inference of switch-like differential expression along single-cell trajectories. *Bioinformatics*, 33(8):1241–1242.

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., and Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., and Shendure, J. (2019a). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496.

Cao, Z.-J., Wei, L., Lu, S., Yang, D.-C., and Gao, G. (2019b). Cell BLAST: Searching large-scale scRNA-seq database via unbiased cell embedding. *bioRxiv*, page 587360.

Cebula, A., Seweryn, M., Rempala, G. A., Pabla, S. S., McIndoe, R. A., Denning, T. L., Bry, L., Kraj, P., Kisielow, P., and Ignatowicz, L. (2013). Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. *Nature*, 497(7448):258–262.

Cepek, K. L., Shaw, S. K., Parker, C. M., Russell, G. J., Morrow, J. S., Rimm, D. L., and Brenner, M. B. (1994). Adhesion between epithelial cells and T lymphocytes mediated by e-cadherin and the αEβ7 integrin. *Nature*, 372(6502):190–193.

Chang, Q., Ornatsky, O. I., Siddiqui, I., Loboda, A., Baranov, V. I., and Hedley, D. W. (2017). Imaging Mass Cytometry. *Cytometry Part A*, 91(2):160–169.

Chen, S., Park, J. H., Rivaud, P., Charles, E., Haliburton, J., Pichiorri, F., and Thomson, M. (2018). Dissecting heterogeneous cell-populations across signaling and disease conditions with PopAlign. *bioRxiv*, page 421354.

Chen, X., Chen, S., and Jiang, R. (2019). EnClaSC: A novel ensemble approach for accurate and robust cell-type classification of single-cell transcriptomes. *bioRxiv*, page 754085.

Chow, Z., Banerjee, A., and Hickey, M. J. (2015). Controlling the fire — tissue-specific mechanisms of effector regulatory t-cell homing. *Immunol. Cell Biol.*, 93(4):355–363.

Cipolletta, D. (2014). Adipose tissue-resident regulatory T cells: phenotypic specialization, functions and therapeutic potential. *Immunology*, 142(4):517–525.

Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., and Reik, W. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9(1):781.

Collins, P. and Billett, F. S. (1995). The terminology of early development: History, concepts, and current usage. *Clinical Anatomy*, 8(6):418–425.

Consortium, T. G. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Coons, A. H., Creech, H. J., and Jones, R. N. (1941). Immunological Properties of an Antibody Containing a Fluorescent Group. *Proceedings of the Society for Experimental Biology and Medicine*, 47(2):200–202.

Cretney, E., Xin, A., Shi, W., Minnich, M., Masson, F., Miasari, M., Belz, G. T., Smyth, G. K., Busslinger, M., Nutt, S. L., and Kallies, A. (2011). The transcription factors blimp-1 and IRF4 jointly control the differentiation and function of effector regulatory T cells. *Nat. Immunol.*, 12(4):304–311.

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., and Shendure, J. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, 174(5):1309–1324.e18.

Damianou, A., Ek, C., Titsias, M., and Lawrence, N. (2012). Manifold relevance determination. *arXiv*.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301.

de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T., and Holstege, F. C. P. (2019). CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research*.

De Simone, M., De Simone, M., Arrigoni, A., Rossetti, G., Gruarin, P., Ranzani, V., Politano, C., Bonnal, R. J. P., Provasi, E., Sarnicola, M. L., Panzeri, I., Moro, M., Crosti, M., Mazzara, S., Vaira, V., Bosari, S., Palleschi, A., Santambrogio, L.,

Bovo, G., Zucchini, N., Totis, M., Gianotti, L., Cesana, G., Perego, R. A., Maroni, N., Ceretti, A. P., Opocher, E., De Francesco, R., Geginat, J., Stunnenberg, H. G., Abrignani, S., and Pagani, M. (2016). Transcriptional landscape of human tissue lymphocytes unveils uniqueness of Tumor-Infiltrating T regulatory cells. *Immunity*, 45(5):1135–1147.

DePasquale, E. A., Dexheimer, P., Schnell, D., Ferchen, K., Hay, S., Valiente-Alandi, I., Blaxall, B. C., Grimes, H. L., and Salomonis, N. (2019). cellHarmony: Cell-level matching and holistic comparison of single-cell transcriptomes. *bioRxiv*, page 412080.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, 278(5338):680–686.

Di Palma, S. and Bodenmiller, B. (2015). Unraveling cell populations in tumors by single-cell mass cytometry. *Current Opinion in Biotechnology*, 31:122–129.

DiSpirito, J. R., Zemmour, D., Ramanan, D., Cho, J., Zilionis, R., Klein, A. M., Benoist, C., and Mathis, D. (2018). Molecular diversification of regulatory T cells in nonlymphoid tissues. *Science Immunology*, 3(27).

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., and Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17.

Domanskyi, S., Szedlak, A., Hawkins, N. T., Wang, J., Paternostro, G., and Piermarocchi, C. (2019). Polled Digital Cell Sorter (p-DCS): Automatic identification of hematological cell types from single cell RNA-sequencing clusters. *BMC Bioinformatics*, 20(1):369.

Dongen, S. V. (2000). Performance Criteria for Graph Clustering and Markov Cluster Experiments. Technical report, NATIONAL RESEARCH INSTITUTE FOR MATHEMATICS AND COMPUTER SCIENCE IN THE.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Bernstein, B. E., Birney, E., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigó, R.,

Hardison, R. C., Hubbard, T. J., Kellis, M., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Khatun, J., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Giddings, M. C., Bernstein, B. E., Epstein, C. B., Shoresh, N., Ernst, J., Kheradpour, P., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Eaton, M. L., Kellis, M., Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H. P., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Risk, B. A., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T. J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T. R., Rosenbloom, K. R., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kent, W. J., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Furey, T. S., Song, L., Grasfeder, L. L., Giresi, P. G., Lee, B.-K., Battenhouse, A., Sheffield, N. C., Simon, J. M., Showers, K. A., Safi, A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Ki Kim, S., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Birney, E., Iyer, V. R., Lieb, J. D., Crawford, G. E., Li, G., Sandhu, K. S., Zheng, M., Wang, P., Luo, O. J., Shahab, A., Fullwood, M. J., Ruan, X., Ruan, Y., Myers, R. M., Pauli, F., Williams, B. A., Gertz, J., Marinov, G. K., Reddy, T. E., Vielmetter, J., Partridge, E., Trout, D., Varley, K. E., Gasper, C., The ENCODE Project Consortium, Overall coordination (data analysis coordination), Data production leads (data production), Lead analysts (data analysis), Writing group, NHGRI project management (scientific management), Principal investigators (steering committee), Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis), Broad Institute Group (data production and analysis), Cold Spring Harbor, Center for Genomic Regulation, B. R. S. I. U. o. L. G. I. o. S. g. d. p. a. a. U. o. G., Data coordination center at UC Santa Cruz (production data coordination), Duke University, University of Texas, A. U. o. N. C.-C. H. g. d. p. a. a. E., Genome Institute of Singapore group (data production and analysis), and HudsonAlpha Institute, UC Irvine, S. g. d. p. a. a.-C. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8):1184–1191.

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2019). CellPhoneDB v2.0: Inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes. *bioRxiv*, page 680926.

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., Doxiadis, G. M., Bontrop, R. E., and Pääbo, S. (2002). Intra- and Interspecific Variation in Primate Gene Expression Patterns. *Science*, 296(5566):340–343.

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjöstedt, E., Lundberg, E., Szigyarto, C. A.-K., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., Nilsson, P., Schwenk, J. M., Lindskog, C., Danielsson, F., Mardinoglu, A., Sivertsson, A., von Feilitzen, K., Forsberg, M., Zwahlen, M., Olsson, I., Navani, S., Huss, M., Nielsen, J., Ponten, F., and Uhlén, M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics*, 13(2):397–406.

Fincher, C. T., Wurtzel, O., Hoog, T. d., Kravarik, K. M., and Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science*, 360(6391):eaaq1736.

Franzén, O., Gan, L.-M., and Björkegren, J. L. M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019.

Gabanyi, I., Muller, P., Feighery, L., Oliveira, T., Costa-Pinto, F., and Mucida, D. (2016). Neuro-immune Interactions Drive Tissue Programming in Intestinal Macrophages. *Cell*, 164(3):378–391.

Gao, X., Hu, D., Gogol, M., and Li, H. (2018). ClusterMap: Comparing analyses across multiple Single Cell RNA-Seq profiles. *bioRxiv*, page 331330.

Geremia, A., Arancibia-Cárcamo, C. V., Fleming, M. P. P., Rust, N., Singh, B., Mortensen, N. J., Travis, S. P. L., and Powrie, F. (2011). IL-23-responsive innate lymphoid cells are increased in inflammatory bowel disease. *J. Exp. Med.*, 208(6):1127–1133.

Gierahn, T. M., Wadsworth Ii, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Love, J. C., and Shalek, A. K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods*, 14(4):395–398.

Glusman, G., Rowen, L., Lee, I., Boysen, C., Roach, J. C., Smit, A. F. A., Wang, K., Koop, B. F., and Hood, L. (2001). Comparative Genomics of the Human and Mouse T Cell Receptor Loci. *Immunity*, 15(3):337–349.

Goldstein, L. D., Chen, Y.-J. J., Dunne, J., Mir, A., Hubschle, H., Guillory, J., Yuan, W., Zhang, J., Stinson, J., Jaiswal, B., Pahuja, K. B., Mann, I., Schaal, T., Chan, L., Anandakrishnan, S., Lin, C.-w., Espinoza, P., Husain, S., Shapiro, H., Swaminathan, K., Wei, S., Srinivasan, M., Seshagiri, S., and Modrusan, Z. (2017). Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics*, 18(1):519.

Golgi, C. and Lipsky, N. G. (1989). On the structure of nerve cells. *Journal of Microscopy*, 155(1):3–7.

Gomez Perdiguero, E., Klapproth, K., Schulz, C., Busch, K., Azzoni, E., Crozet, L., Garner, H., Trouillet, C., de Bruijn, M. F., Geissmann, F., and Rodewald, H.-R. (2015). Tissue-resident macrophages originate from yolk-sac-derived erythro-myeloid progenitors. *Nature*, 518(7540):547–551.

Gordon, S. and Martinez-Pomares, L. (2017). Physiological roles of macrophages. *Pflügers Archiv - European Journal of Physiology*, 469(3):365–374.

Gorin, G., Svensson, V., and Pachter, L. (2019). RNA velocity and protein acceleration from single-cell multiomics experiments. *bioRxiv*, page 658401.

Gosselin, D., Link, V. M., Romanoski, C., Fonseca, G., Eichenfield, D., Spann, N., Stender, J., Chun, H., Garner, H., Geissmann, F., and Glass, C. (2014). Environment Drives Selection and Function of Enhancers Controlling Tissue-Specific Macrophage Identities. *Cell*, 159(6):1327–1340.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

Guo, J., Grow, E. J., Mlcochova, H., Maher, G. J., Lindskog, C., Nie, X., Guo, Y., Takei, Y., Yun, J., Cai, L., Kim, R., Carrell, D. T., Goriely, A., Hotaling, J. M., and Cairns, B. R. (2018). The adult human testis transcriptional cell atlas. *Cell Research*, 28(12):1141–1157.

Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E., and Tilgner, H. U. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology*, 36(12):1197–1202.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S. R., Aguet, F., Gelfand, E., Ardlie, K., Weitz, D. A., Rozenblatt-Rosen, O., Zhang, F., and Regev, A. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods*, 14(10):955–958.

Hagai, T., Chen, X., Miragaia, R. J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.-E., Proserpio, V., Donati, G., Bossini-Castillo, L., Braga, F. A. V., Naamati, G., Fletcher, J., Stephenson, E., Vegh, P., Trynka, G., Kondova, I., Dennis, M., Haniffa, M., Nourmohammad, A., Lässig, M., and Teichmann, S. A. (2018). Gene expression variability across cells and species shapes innate immunity. *Nature*, 563(7730):197.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427.

Halle, S., Halle, O., and Förster, R. (2017). Mechanisms and Dynamics of T Cell-Mediated Cytotoxicity In Vivo. *Trends in Immunology*, 38(6):432–443.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G.-C., Chen, M., and Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 172(5):1091–1107.e17.

Haribhai, D., Lin, W., Relland, L. M., Truong, N., Williams, C. B., and Chatila, T. A. (2007). Regulatory T cells dynamically control the primary immune response to foreign antigen. *J. Immunol.*, 178(5):2961–2972.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K. J., Rozenblatt-Rosen, O., Dor, Y., Regev, A., and Yanai, I. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17(1):77.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, 2(3):666–673.

Hayashi, T., Ozaki, H., Sasagawa, Y., Umeda, M., Danno, H., and Nikaido, I. (2018). Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nature Communications*, 9(1):1–16.

Henry, G. H., Malewska, A., Joseph, D. B., Malladi, V. S., Lee, J., Torrealba, J., Mauck, R. J., Gahan, J. C., Raj, G. V., Roehrborn, C. G., Hon, G. C., MacConmara, M. P., Reese, J. C., Hutchinson, R. C., Vezina, C. M., and Strand, D. W. (2018). A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. *Cell Reports*, 25(12):3530–3542.e5.

Hie, B., Bryson, B., and Berger, B. (2019a). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(6):685–691.

Hie, B., Cho, H., DeMeo, B., Bryson, B., and Berger, B. (2019b). Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. *Cell Systems*, 8(6):483–493.e7.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., Akbani, R., Bowlby, R., Wong, C. K., Wiznerowicz, M., Sanchez-Vega, F., Robertson, A. G., Schneider, B. G., Lawrence, M. S., Noushmehr, H., Malta, T. M., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., Ferguson, M. L., Hutter, C. M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Zhang, J. J., Chudamani, S., Liu, J., Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Cho, J., DeFreitas, T., Frazer, S., Gehlenborg, N., Getz, G., Heiman, D. I., Kim, J., Lawrence, M. S., Lin, P., Meier, S., Noble, M. S., Saksena, G., Voet, D., Zhang, H., Bernard, B., Chambwe, N., Dhankani,

V., Knijnenburg, T., Kramer, R., Leinonen, K., Liu, Y., Miller, M., Reynolds, S., Shmulevich, I., Thorsson, V., Zhang, W., Akbani, R., Broom, B. M., Hegde, A. M., Ju, Z., Kanchi, R. S., Korkut, A., Li, J., Liang, H., Ling, S., Liu, W., Lu, Y., Mills, G. B., Ng, K.-S., Rao, A., Ryan, M., Wang, J., Weinstein, J. N., Zhang, J., Abeshouse, A., Armenia, J., Chakravarty, D., Chatila, W. K., Bruijn, I. d., Gao, J., Gross, B. E., Heins, Z. J., Kundra, R., La, K., Ladanyi, M., Luna, A., Nissan, M. G., Ochoa, A., Phillips, S. M., Reznik, E., Sanchez-Vega, F., Sander, C., Schultz, N., Sheridan, R., Sumer, S. O., Sun, Y., Taylor, B. S., Wang, J., Zhang, H., Anur, P., Peto, M., Spellman, P., Benz, C., Stuart, J. M., Wong, C. K., Yau, C., Hayes, D. N., Parker, J. S., Wilkerson, M. D., Ally, A., Balasundaram, M., Bowlby, R., Brooks, D., Carlsen, R., Chuah, E., Dhalla, N., Holt, R., Jones, S. J. M., Kasaian, K., Lee, D., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Mungall, K., Robertson, A. G., Sadeghi, S., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Tse, K., Wong, T., Berger, A. C., Beroukhim, R., Cherniack, A. D., Cibulskis, C., Gabriel, S. B., Gao, G. F., Ha, G., Meyerson, M., Schumacher, S. E., Shih, J., Kucherlapati, M. H., Kucherlapati, R. S., Baylin, S., Cope, L., Danilova, L., Bootwalla, M. S., Lai, P. H., Maglinte, D. T., Berg, D. J. V. D., Weisenberger, D. J., Auman, J. T., Balu, S., Bodenheimer, T., Fan, C., Hoadley, K. A., Hoyle, A. P., Jefferys, S. R., Jones, C. D., Meng, S., Mieczkowski, P. A., Mose, L. E., Perou, A. H., Perou, C. M., Roach, J., Shi, Y., Simons, J. V., Skelly, T., Soloway, M. G., Tan, D., Veluvolu, U., Fan, H., Hinoue, T., Laird, P. W., Shen, H., Zhou, W., Bellair, M., Chang, K., Covington, K., Creighton, C. J., Dinh, H., Doddapaneni, H., Donehower, L. A., Drummond, J., Gibbs, R. A., Glenn, R., Hale, W., Han, Y., Hu, J., Korchina, V., Lee, S., Lewis, L., Li, W., Liu, X., Morgan, M., Morton, D., Muzny, D., Santibanez, J., Sheth, M., Shinbrot, E., Wang, L., Wang, M., Wheeler, D. A., Xi, L., Zhao, F., Hess, J., Appelbaum, E. L., Bailey, M., Cordes, M. G., Ding, L., Fronick, C. C., Fulton, L. A., Fulton, R. S., Kandoth, C., Mardis, E. R., McLellan, M. D., Miller, C. A., Schmidt, H. K., Wilson, R. K., Crain, D., Curley, E., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Sherman, M., Thompson, E., Yena, P., Bowen, J., Gastier-Foster, J. M., Gerken, M., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Wise, L., Zmuda, E., Corcoran, N., Costello, T., Hovens, C., Carvalho, A. L., Carvalho, A. C. d., Fregnani, J. H., Longatto-Filho, A., Reis, R. M., Scapulatempo-Neto, C., Silveira, H. C. S., Vidal, D. O., Burnette, A., Eschbacher, J., Hermes, B., Noss, A., Singh, R., Anderson, M. L., Castro, P. D., Ittmann, M., Huntsman, D., Kohl, B., Le, X., Thorp, R., Andry, C., Duffy, E. R., Lyadov, V., Paklina, O., Setdikova, G., Shabunin, A., Tavobilov, M., McPherson, C., Warnick, R., Berkowitz, R., Cramer, D., Feltmate, C., Horowitz, N., Kibel, A., Muto, M., Raut, C. P., Malykh, A., Barnholtz-Sloan, J. S., Barrett, W., Devine, K., Fulop, J., Ostrom, Q. T., Shimmel, K., Wolinsky, Y., Sloan, A. E., Rose, A. D., Giuliante, F., Goodman, M., Karlan, B. Y., Hagedorn, C. H., Eckman, J., Harr, J., Myers, J., Tucker, K., Zach, L. A., Deyarmin, B., Hu, H., Kvecher, L., Larson, C., Mural, R. J., Somiari, S., Vicha, A., Zelinka, T., Bennett, J., Iacocca, M., Rabeno, B., Swanson, P., Latour, M., Lacombe, L., Têtu, B., Bergeron, A., McGraw, M., Staugaitis, S. M., Chabot, J., Hibshoosh, H., Sepulveda, A., Su, T., Wang, T., Potapova, O., Voronina, O., Desjardins, L., Mariani, O., Roman-Roman, S., Sastre, X., Stern, M.-H., Cheng, F., Signoretti, S., Berchuck, A., Bigner, D., Lipp, E., Marks, J., McCall, S., McLendon, R., Secord, A., Sharp, A., Behera, M., Brat, D. J., Chen, A., Delman, K., Force, S., Khuri, F., Magliocca, K., Maithel, S., Olson, J. J., Owonikoko, T., Pickens, A., Ramalingam, S., Shin, D. M., Sica, G., Meir, E. G. V., Zhang, H.,

Eijckenboom, W., Gillis, A., Korpershoek, E., Looijenga, L., Oosterhuis, W., Stoop, H., Kessel, K. E. v., Zwarthoff, E. C., Calatozzolo, C., Cuppini, L., Cuzzubbo, S., DiMeco, F., Finocchiaro, G., Mattei, L., Perin, A., Pollo, B., Chen, C., Houck, J., Lohavanichbutr, P., Hartmann, A., Stoehr, C., Stoehr, R., Taubert, H., Wach, S., Wullich, B., Kycler, W., Murawa, D., Wiznerowicz, M., Chung, K., Edenfield, W. J., Martin, J., Baudin, E., Bubley, G., Bueno, R., Rienzo, A. D., Richards, W. G., Kalkanis, S., Mikkelsen, T., Noushmehr, H., Scarpace, L., Girard, N., Aymerich, M., Campo, E., Giné, E., Guillermo, A. L., Bang, N. V., Hanh, P. T., Phu, B. D., Tang, Y., Colman, H., Evason, K., Dottino, P. R., Martignetti, J. A., Gabra, H., Juhl, H., Akeredolu, T., Stepa, S., Hoon, D., Ahn, K., Kang, K. J., Beuschlein, F., Breggia, A., Birrer, M., Bell, D., Borad, M., Bryce, A. H., Castle, E., Chandan, V., Cheville, J., Copland, J. A., Farnell, M., Flotte, T., Giama, N., Ho, T., Kendrick, M., Kocher, J.-P., Kopp, K., Moser, C., Nagorney, D., O'Brien, D., O'Neill, B. P., Patel, T., Petersen, G., Que, F., Rivera, M., Roberts, L., Smallridge, R., Smyrk, T., Stanton, M., Thompson, R. H., Torbenson, M., Yang, J. D., Zhang, L., Brimo, F., Ajani, J. A., Gonzalez, A. M. A., Behrens, C., Bondaruk, O., Broaddus, R., Czerniak, B., Esmaeli, B., Fujimoto, J., Gershenwald, J., Guo, C., Lazar, A. J., Logothetis, C., Meric-Bernstam, F., Moran, C., Ramondetta, L., Rice, D., Sood, A., Tamboli, P., Thompson, T., Troncoso, P., Tsao, A., Wistuba, I., Carter, C., Haydu, L., Hersey, P., Jakrot, V., Kakavand, H., Kefford, R., Lee, K., Long, G., Mann, G., Quinn, M., Saw, R., Scolyer, R., Shannon, K., Spillane, A., Stretch, J., Synott, M., Thompson, J., Wilmott, J., Al-Ahmadie, H., Chan, T. A., Ghossein, R., Gopalan, A., Levine, D. A., Reuter, V., Singer, S., Singh, B., Tien, N. V., Broudy, T., Mirsaidi, C., Nair, P., Drwiega, P., Miller, J., Smith, J., Zaren, H., Park, J.-W., Hung, N. P., Kebebew, E., Linehan, W. M., Metwalli, A. R., Pacak, K., Pinto, P. A., Schiffman, M., Schmidt, L. S., Vocke, C. D., Wentzensen, N., Worrell, R., Yang, H., Moncrieff, M., Goparaju, C., Melamed, J., Pass, H., Botnariuc, N., Caraman, I., Cernat, M., Chemencedji, I., Clipca, A., Doruc, S., Gorincioi, G., Mura, S., Pirtac, M., Stancul, I., Tcaciuc, D., Albert, M., Alexopoulou, I., Arnaout, A., Bartlett, J., Engel, J., Gilbert, S., Parfitt, J., Sekhon, H., Thomas, G., Rassl, D. M., Rintoul, R. C., Bifulco, C., Tamakawa, R., Urba, W., Hayward, N., Timmers, H., Antenucci, A., Facciolo, F., Grazi, G., Marino, M., Merola, R., Krijger, R. d., Gimenez-Roqueplo, A.-P., Piché, A., Chevalier, S., McKercher, G., Birsoy, K., Barnett, G., Brewer, C., Farver, C., Naska, T., Pennell, N. A., Raymond, D., Schilero, C., Smolenski, K., Williams, F., Morrison, C., Borgia, J. A., Liptay, M. J., Pool, M., Seder, C. W., Junker, K., Omberg, L., Dinkin, M., Manikhas, G., Alvaro, D., Bragazzi, M. C., Cardinale, V., Carpino, G., Gaudio, E., Chesla, D., Cottingham, S., Dubina, M., Moiseenko, F., Dhanasekaran, R., Becker, K.-F., Janssen, K.-P., Slotta-Huspenina, J., Abdel-Rahman, M. H., Aziz, D., Bell, S., Cebulla, C. M., Davis, A., Duell, R., Elder, J. B., Hilty, J., Kumar, B., Lang, J., Lehman, N. L., Mandt, R., Nguyen, P., Pilarski, R., Rai, K., Schoenfield, L., Senecal, K., Wakely, P., Hansen, P., Lechan, R., Powers, J., Tischler, A., Grizzle, W. E., Sexton, K. C., Kastl, A., Henderson, J., Porten, S., Waldmann, J., Fassnacht, M., Asa, S. L., Schadendorf, D., Couce, M., Graefen, M., Huland, H., Sauter, G., Schlomm, T., Simon, R., Tennstedt, P., Olabode, O., Nelson, M., Bathe, O., Carroll, P. R., Chan, J. M., Disaia, P., Glenn, P., Kelley, R. K., Landen, C. N., Phillips, J., Prados, M., Simko, J., Smith-McCune, K., VandenBerg, S., Roggin, K., Fehrenbach, A., Kendler, A., Sifri, S., Steele, R., Jimeno, A., Carey, F., Forgie, I., Mannelli, M., Carney, M., Hernandez, B., Campos, B., Herold-Mende, C., Jungk, C., Unterberg, A., Deimling, A. v., Bossler, A., Galbraith, J., Jacobus, L., Knudson,

M., Knutson, T., Ma, D., Milhem, M., Sigmund, R., Godwin, A. K., Madan, R., Rosenthal, H. G., Adebamowo, C., Adebamowo, S. N., Boussioutas, A., Beer, D., Giordano, T., Mes-Masson, A.-M., Saad, F., Bocklage, T., Landrum, L., Mannel, R., Moore, K., Moxley, K., Postier, R., Walker, J., Zuna, R., Feldman, M., Valdivieso, F., Dhir, R., Luketich, J., Pinero, E. M. M., Quintero-Aguilo, M., Carlotti, C. G., Santos, J. S. D., Kemp, R., Sankarankuty, A., Tirapelli, D., Catto, J., Agnew, K., Swisher, E., Creaney, J., Robinson, B., Shelley, C. S., Godwin, E. M., Kendall, S., Shipman, C., Bradford, C., Carey, T., Haddad, A., Moyer, J., Peterson, L., Prince, M., Rozek, L., Wolf, G., Bowman, R., Fong, K. M., Yang, I., Korst, R., Rathmell, W. K., Fantacone-Campbell, J. L., Hooke, J. A., Kovatich, A. J., Shriver, C. D., DiPersio, J., Drake, B., Govindan, R., Heath, S., Ley, T., Tine, B. V., Westervelt, P., Rubin, M. A., Lee, J. I., Aredes, N. D., Mariamidze, A., Stuart, J. M., Benz, C. C., and Laird, P. W. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):291–304.e6.

Hooke, R. (1667). *Micrographia: Or, Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses, with Observations and Inquiries Thereupon*. History of microscopy series. Science Heritage.

Hori, S., Nomura, T., and Sakaguchi, S. (2003). Control of Regulatory T Cell Development by the Transcription Factor Foxp3. *Science*, 299(5609):1057–1061.

Hou, R., Denisenko, E., and Forrest, A. R. R. (2019). scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics*.

Hoyer, F. F., Naxerova, K., Schloss, M. J., Hulsmans, M., Nair, A. V., Dutta, P., Calcagno, D. M., Herisson, F., Anzai, A., Sun, Y., Wojtkiewicz, G., Rohde, D., Frodermann, V., Vandoorne, K., Courties, G., Iwamoto, Y., Garris, C. S., Williams, D. L., Breton, S., Brown, D., Whalen, M., Libby, P., Pittet, M. J., King, K. R., Weissleder, R., Swirski, F. K., and Nahrendorf, M. (2019). Tissue-Specific Macrophage Responses to Remote Injury Impact the Outcome of Subsequent Local Immune Challenge. *Immunity*, 51(5):899–914.e7.

Hu, H., Miao, Y.-R., Jia, L.-H., Yu, Q.-Y., Zhang, Q., and Guo, A.-Y. (2019). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Research*, 47(D1):D33–D38.

Hu, Z.-Q. and Zhao, W.-H. (2015). The IL-33/ST2 axis is specifically required for development of adipose tissue-resident regulatory T cells. *Cell. Mol. Immunol.*, 12(5):521–524.

Huehn, J., Siegmund, K., Lehmann, J. C. U., Siewert, C., Haubold, U., Feuerer, M., Debes, G. F., Lauber, J., Frey, O., Przybylski, G. K., Niesner, U., Rosa, M. d. l., Schmidt, C. A., Bräuer, R., Buer, J., Scheffold, A., and Hamann, A. (2004). Developmental Stage, Phenotype, and Migration Distinguish Naive- and Effector/Memory-like CD4+ Regulatory T Cells. *Journal of Experimental Medicine*, 199(3):303–313.

Hughes, T. K., Wadsworth, M. H., Gierahn, T. M., Do, T., Weiss, D., Andrade, P. R., Ma, F., Silva, B. J. d. A., Shao, S., Tsoi, L. C., Ordovas-Montanes, J., Gudjonsson, J. E.,

Modlin, R. L., Love, J. C., and Shalek, A. K. (2019). Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology. *bioRxiv*, page 689273.

Ikebuchi, R., Teraguchi, S., Vandenbon, A., Honda, T., Shand, F. H. W., Nakanishi, Y., Watanabe, T., and Tomura, M. (2016). A rare subset of skin-tropic regulatory T cells expressing Il10/Gzmb inhibits the cutaneous immune response. *Sci. Rep.*, 6:35002.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7):1160–1167.

Ivanov, I. I., Frutos, R. d. L., Manel, N., Yoshinaga, K., Rifkin, D. B., Sartor, R. B., Finlay, B. B., and Littman, D. R. (2008). Specific microbiota direct the differentiation of IL-17-producing t-helper cells in the mucosa of the small intestine. *Cell Host Microbe*, 4(4):337–349.

Izcue, A., Coombes, J. L., and Powrie, F. (2009). Regulatory lymphocytes and intestinal inflammation. *Annu. Rev. Immunol.*, 27(1):313–338.

Jaitin, D. A., Adlung, L., Thaiss, C. A., Weiner, A., Li, B., Descamps, H., Lundgren, P., Bleriot, C., Liu, Z., Deczkowska, A., Keren-Shaul, H., David, E., Zmora, N., Eldar, S. M., Lubezky, N., Shibolet, O., Hill, D. A., Lazar, M. A., Colonna, M., Ginhoux, F., Shapiro, H., Elinav, E., and Amit, I. (2019). Lipid-Associated Macrophages Control Metabolic Homeostasis in a Trem2-Dependent Manner. *Cell*, 178(3):686–698.e14.

James, K. R., Gomes, T., Elmentaite, R., Kumar, N., Gulliver, E. L., King, H. W., Stares, M. D., Bareham, B. R., Ferdinand, J. R., Petrova, V. N., Polanski, K., Forster, S. C., Jarvis, L. B., Suchanek, O., Howlett, S., James, L. K., Jones, J. L., Meyer, K. B., Clatworthy, M. R., Saeb-Parsy, K., Lawley, T. D., and Teichmann, S. A. (2019). Distinct microbial and immune niches of the human colon. *bioRxiv*, page 2019.12.12.871657.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127.

Josefowicz, S. Z., Lu, L.-F., and Rudensky, A. Y. (2012). Regulatory T cells: mechanisms of differentiation and function. *Annu. Rev. Immunol.*, 30:531–564.

Julius, M. H., Masuda, T., and Herzenberg, L. A. (1972). Demonstration That Antigen-Binding Cells Are Precursors of Antibody-Producing Cells After Purification with a Fluorescence-Activated Cell Sorter. *Proceedings of the National Academy of Sciences of the United States of America*, 69(7):1934–1938.

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1):1930.

Keren-Shaul, H., Kenigsberg, E., Jaitin, D. A., David, E., Paul, F., Tanay, A., and Amit, I. (2019). MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing. *Nature Protocols*, 14(6):1841–1862.

Kim, S. V., Xiang, W. V., Kwak, C., Yang, Y., Lin, X. W., Ota, M., Sarpel, U., Rifkin, D. B., Xu, R., and Littman, D. R. (2013). GPR15-mediated homing controls immune homeostasis in the large intestine mucosa. *Science*, 340(6139):1456–1459.

Kimpton, W. G., Washington, E. A., Cahill, R. N. P., and Miyasaka, M. (1995). Virgin $\alpha\beta$ and $\gamma\delta$ T cells recirculate extensively through peripheral tissues and skin during normal development of the fetal immune system. *Int. Immunol.*, 7(10):1567–1577.

Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods*, 15(5):359–362.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5):1187–1201.

Korsunsky, I., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2018). Fast, sensitive, and accurate integration of single cell data with Harmony. *bioRxiv*, page 461954.

Krangel, M. S. (2009). Mechanics of T cell receptor gene rearrangement. *Current Opinion in Immunology*, 21(2):133–139.

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2017). A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, 18(2):205–214.

Kumar, B. V., Connors, T. J., and Farber, D. L. (2018). Human T Cell Development, Localization, and Function throughout Life. *Immunity*, 48(2):202–213.

Köhler, N. D., Büttner, M., and Theis, F. J. (2019). Deep learning does not outperform classical machine learning for cell-type annotation. *bioRxiv*, page 653907.

La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L. E., Stott, S. R. W., Toledo, E. M., Villaescusa, J. C., Lönnerberg, P., Ryge, J., Barker, R. A., Arenas, E., and Linnarsson, S. (2016). Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell*, 167(2):566–580.e19.

Laurent, P., Jolivel, V., Manicki, P., Chiu, L., Contin-Bordes, C., Truchetet, M.-E., and Pradeu, T. (2017). Immune-Mediated Repair: A Matter of Plasticity. *Frontiers in Immunology*, 8.

Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S., and Amit, I. (2014). Tissue-Resident Macrophage Enhancer Landscapes Are Shaped by the Local Microenvironment. *Cell*, 159(6):1312–1326.

Leeuwenhoeck M (1674). Microscopical observations from Leeuwenhoeck, concerning blood, milk, bones, the brain, spitle, and cuticula, &c. communicated by the said observer to the Publisher in a letter, dated June 1. 1674. *Philosophical Transactions of the Royal Society of London*, 9(106):121–131.

Leeuwenhoek Antoni Van (1677). Observationes D. Anthonii Lewenhoeck, de natis'e semine genitali animalculis. *Philosophical Transactions of the Royal Society of London*, 12(142):1040–1046.

Li, C., DiSpirito, J. R., Zemmour, D., Spallanzani, R. G., Kuswanto, W., Benoist, C., and Mathis, D. (2018a). TCR transgenic mice reveal stepwise, multi-site acquisition of the distinctive Fat-Treg phenotype. *Cell*, 174(2):285–299.e12.

Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Ren, X., and Zhang, Z. (2019a). SciBet: a fast classifier for cell type identification using single cell RNA sequencing data. *bioRxiv*, page 645358.

Li, J., Tan, J., Martino, M. M., and Lui, K. O. (2018b). Regulatory T-Cells: Potential Regulator of Tissue Repair and Regeneration. *Frontiers in Immunology*, 9.

Li, J. J., Huang, H., Bickel, P. J., and Brenner, S. E. (2014). Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Research*, 24(7):1086–1101.

Li, N., Unen, V. v., Abdelaal, T., Guo, N., Kasatskaya, S. A., Ladell, K., McLaren, J. E., Egorov, E. S., Izraelson, M., Lopes, S. M. C. d. S., Höllt, T., Britanova, O. V., Eggermont, J., Miranda, N. F. C. C. d., Chudakov, D. M., Price, D. A., Lelieveldt, B. P. F., and Koning, F. (2019b). Memory CD4 + T cells are generated in the human fetal intestine. *Nature Immunology*, 20(3):301–312.

Lieberman, Y., Rokach, L., and Shay, T. (2018). CaSTLe – Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLOS ONE*, 13(10):e0205499.

Lin, Y., Cao, Y., Kim, H. J., Salim, A., Speed, T. P., Lin, D., Yang, P., and Yang, J. Y. H. (2019). scClassify: hierarchical classification of cells. *bioRxiv*, page 776948.

Lindeman, I., Emerton, G., Mamanova, L., Snir, O., Polanski, K., Qiao, S.-W., Sollid, L. M., Teichmann, S. A., and Stubbington, M. J. T. (2018). BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nature Methods*, 15(8):563–565.

Liston, A. and Gray, D. H. D. (2014). Homeostatic control of regulatory T cell diversity. *Nature Reviews Immunology*, 14(3):154–165.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N. J., Nicolae, D. L., Gamazon, E. R., Im, H. K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E. T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson,

J. M., Wilder, E. L., Derr, L. K., Green, E. D., Struewing, J. P., Temple, G., Volpi, S., Boyer, J. T., Thomson, E. J., Guyer, M. S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T. R., Koester, S. E., Little, A. R., Bender, P. K., Lehner, T., Yao, Y., Compton, C. C., Vaught, J. B., Sawyer, S., Lockhart, N. C., Demchok, J., and Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45:580–585.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053.

Lotfollahi, M., Wolf, F. A., and Theis, F. J. (2018). Generative modeling and latent space arithmetics predict single-cell perturbation response across cell types, studies and species. *bioRxiv*, page 478503.

Luckheeram, R. V., Zhou, R., Verma, A. D., and Xia, B. (2012). CD4+T Cells: Differentiation and Functions. *Clinical and Developmental Immunology*, 2012.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746.

Lönnberg, T., Svensson, V., James, K. R., Fernandez-Ruiz, D., Sebina, I., Montandon, R., Soon, M. S. F., Fogg, L. G., Nair, A. S., Liligeto, U. N., Stubbington, M. J. T., Ly, L.-H., Bagger, F. O., Zwiessele, M., Lawrence, N. D., Souza-Fonseca-Guimaraes, F., Bunn, P. T., Engwerda, C. R., Heath, W. R., Billker, O., Stegle, O., Haque, A., and Teichmann, S. A. (2017). Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria. *Science Immunology*, 2(9):eaal2192.

Ma, F. and Pellegrini, M. (2019). Automated identification of Cell Types in Single Cell RNA Sequencing. *bioRxiv*, page 532093.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214.

Madissoon, E., Wilbrey-Clark, A., Miragaia, R. J., Saeb-Parsy, K., Mahbubani, K., Georgakopoulos, N., Harding, P., Polanski, K., Nowicki-Osuch, K., Fitzgerald, R. C., Loudon, K. W., Ferdinand, J. R., Clatworthy, M. R., Tsingene, A., Dongen, S. V., Dabrowska, M., Patel, M., Stubbington, M. J. T., Teichmann, S., Stegle, O., and Meyer, K. B. (2019). Lung, spleen and oesophagus tissue remains stable for scRNAseq in cold preservation. *bioRxiv*, page 741405.

Malhotra, N., Leyva-Castillo, J. M., Jadhav, U., Barreiro, O., Kam, C., O'Neill, N. K., Meylan, F., Chambon, P., von Andrian, U. H., Siegel, R. M., Wang, E. C., Shivdasani, R., and Geha, R. S. (2018). Rorα-expressing T regulatory cells restrain allergic skin inflammation. *Science Immunology*, 3(21):eaao6923.

Malik, B. T., Byrne, K. T., Vella, J. L., Zhang, P., Shabaneh, T. B., Steinberg, S. M., Molodtsov, A. K., Bowers, J. S., Angeles, C. V., Paulos, C. M., Huang, Y. H., and Turk, M. J. (2017). Resident memory T cells in the skin mediate durable immunity to melanoma. *Sci Immunol*, 2(10).

Manno, G. L., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., Bruggen, D. v., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, 560(7719):494.

Masuda, T., Sankowski, R., Staszewski, O., Böttcher, C., Amann, L., Sagar, Scheiwe, C., Nessler, S., Kunz, P., Loo, G. v., Coenen, V. A., Reinacher, P. C., Michel, A., Sure, U., Gold, R., Grün, D., Priller, J., Stadelmann, C., and Prinz, M. (2019). Spatial and temporal heterogeneity of mouse and human microglia at single-cell resolution. *Nature*, 566(7744):388–392.

Matsushima, H. and Takashima, A. (2010). Bidirectional homing of tregs between the skin and lymph nodes. *J. Clin. Invest.*, 120(3):653–656.

Mazzarello, P. (1999). A unifying concept: the history of cell theory. *Nature Cell Biology*, 1(1):E13.

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*. arXiv: 1802.03426.

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., Consortium, T. G., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G., and Guigó, R. (2015). The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665.

Menon, R., Otto, E. A., Kokoruda, A., Zhou, J., Zhang, Z., Yoon, E., Chen, Y.-C., Troyanskaya, O., Spence, J. R., Kretzler, M., and Cebrián, C. (2018). Single-cell analysis of progenitor cell dynamics and lineage specification in the human fetal kidney. *Development*, 145(16):dev164038.

Mereu, E., Iacono, G., Guillaumet-Adkins, A., Moutinho, C., Lunazzi, G., Santos, C., Miguel-Escalada, I., Ferrer, J., Real, F. X., Gut, I., and Heyn, H. (2018). matchSCore: Matching Single-Cell Phenotypes Across Tools and Experiments. *bioRxiv*, page 314831.

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., MacCarthy, D. J., Alvarez, A., Batlle, E., Sagar, Grün, D., Lau, J. K., Boutet, S. C., Sanada, C., Ooi, A., Jones, R. C., Kaihara, K., Brampton, C., Talaga, Y., Sasagawa, Y., Tanaka, K., Hayashi, T., Nikaido, I., Fischer, C., Sauer, S., Trefzer, T., Conrad, C., Adiconis, X., Nguyen, L. T., Regev, A., Levin, J. Z., Parekh, S., Janjic, A., Wange, L. E., Bagnoli, J. W., Enard, W., Gut, M., Sandberg, R., Gut, I., Stegle, O., and Heyn, H. (2019). Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects. *bioRxiv*, page 630087.

Miragaia, R. J., Gomes, T., Chomka, A., Jardine, L., Riedel, A., Hegazy, A. N., Whibley, N., Tucci, A., Chen, X., Lindeman, I., Emerton, G., Krausgruber, T., Shields, J., Haniffa, M., Powrie, F., and Teichmann, S. A. (2019). Single-Cell Transcriptomics of Regulatory T Cells Reveals Trajectories of Tissue Adaptation. *Immunity*, 50(2):493–504.e7.

Montoro, D. T., Haber, A. L., Biton, M., Vinarsky, V., Lin, B., Birket, S. E., Yuan, F., Chen, S., Leung, H. M., Villoria, J., Rogel, N., Burgin, G., Tsankov, A. M., Waghray, A., Slyper, M., Waldman, J., Nguyen, L., Dionne, D., Rozenblatt-Rosen, O., Tata, P. R., Mou, H., Shivaraju, M., Bihler, H., Mense, M., Tearney, G. J., Rowe, S. M., Engelhardt, J. F., Regev, A., and Rajagopal, J. (2018). A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature*, 560(7718):319.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628.

Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A., and Coffman, R. L. (1986). Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *The Journal of Immunology*, 136(7):2348–2357.

Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M. A., Carlotti, F., de Koning, E. J. P., and van Oudenaarden, A. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, 3(4):385–394.e3.

Nagar, M., Jacob-Hirsch, J., Vernitsky, H., Berkun, Y., Ben-Horin, S., Amariglio, N., Bank, I., Kloog, Y., Rechavi, G., and Goldstein, I. (2010). TNF Activates a NF-kB–Regulated Cellular Program in Human CD45ra $^-$ Regulatory T Cells that Modulates Their Suppressive Function. *The Journal of Immunology*, 184(7):3570–3581.

Nitta, N., Sugimura, T., Isozaki, A., Mikami, H., Hiraki, K., Sakuma, S., Iino, T., Arai, F., Endo, T., Fujiwaki, Y., Fukuzawa, H., Hase, M., Hayakawa, T., Hiramatsu, K., Hoshino, Y., Inaba, M., Ito, T., Karakawa, H., Kasai, Y., Koizumi, K., Lee, S., Lei, C., Li, M., Maeno, T., Matsusaka, S., Murakami, D., Nakagawa, A., Oguchi, Y., Oikawa, M., Ota, T., Shiba, K., Shintaku, H., Shirasaki, Y., Suga, K., Suzuki, Y., Suzuki, N., Tanaka, Y., Tezuka, H., Toyokawa, C., Yalikun, Y., Yamada, M., Yamagishi, M., Yamano, T., Yasumoto, A., Yatomi, Y., Yazawa, M., Di Carlo, D., Hosokawa, Y., Uemura, S., Ozeki, Y., and Goda, K. (2018). Intelligent Image-Activated Cell Sorting. *Cell*, 175(1):266–276.e13.

Nowakowski, T. J., Bhaduri, A., Pollen, A. A., Alvarado, B., Mostajo-Radji, M. A., Lullo, E. D., Haeussler, M., Sandoval-Espinosa, C., Liu, S. J., Velmeshev, D., Ounadjela, J. R., Shuga, J., Wang, X., Lim, D. A., West, J. A., Leyrat, A. A., Kent, W. J., and Kriegstein, A. R. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science*, 358(6368):1318–1323.

Ohnmacht, C., Park, J.-H., Cording, S., Wing, J. B., Atarashi, K., Obata, Y., Gaboriau-Routhiau, V., Marques, R., Dulauroy, S., Fedoseeva, M., Busslinger, M., Cerf-Bensussan, N., Boneca, I. G., Voehringer, D., Hase, K., Honda, K., Sakaguchi, S., and Eberl, G. (2015). MUCOSAL IMMUNOLOGY. the microbiota regulates type 2 immunity through RORγt^+ T cells. *Science*, 349(6251):989–993.

Panduro, M., Benoist, C., and Mathis, D. (2016). Tissue tregs. *Annu. Rev. Immunol.*, 34:609–633.

Park, J.-E., Polański, K., Meyer, K., and Teichmann, S. A. (2018). Fast Batch Alignment of Single Cell Transcriptomes Unifies Multiple Mouse Cell Atlases into an Integrated Landscape. *bioRxiv*, page 397042.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Picelli, S., Faridani, O. R., Bjorklund, A. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1):171–181.

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F. J., Kocks, C., and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391):eaaq1723.

Pliner, H. A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. *bioRxiv*, page 538652.

Plitas, G., Konopacki, C., Wu, K., Bos, P. D., Morrow, M., Putintseva, E. V., Chudakov, D. M., and Rudensky, A. Y. (2016). Regulatory T cells exhibit distinct features in human breast cancer. *Immunity*, 45(5):1122–1134.

Polański, K., Park, J. E., Young, M. D., Miao, Z., Meyer, K. B., and Teichmann, S. A. (2019). BBKNN: Fast Batch Alignment of Single Cell Transcriptomes. *Bioinformatics (Oxford, England)*.

Popescu, D.-M., Botting, R. A., Stephenson, E., Green, K., Jardine, L., Calderbank, E. F., Efremova, M., Acres, M., Maunder, D., Vegh, P., Goh, I., Gitton, Y., Park, J., Polanski, K., Vento-Tormo, R., Miao, Z., Rowell, R., McDonald, D., Fletcher, J., Dixon, D., Poyner, E., Reynolds, G., Mather, M., Moldovan, C., Mamanova, L., Greig, F., Young, M., Meyer, K., Lisgo, S., Bacardit, J., Fuller, A., Millar, B., Innes, B., Lindsay, S., Stubbington, M. J. T., Kowalczyk, M. S., Li, B., Ashenbrg, O., Tabaka, M., Dionne, D., Tickle, T. L., Slyper, M., Rozenblatt-Rosen, O., Filby, A., Villani, A.-C., Roy, A., Regev, A., Chedotal, A., Roberts, I., Göttgens, B., Laurenti, E., Behjati, S., Teichmann, S. A., and Haniffa, M. (2019). Decoding the development of the blood and immune systems during human fetal liver haematopoiesis. *bioRxiv*, page 654210.

Prasad, A. and Alizadeh, E. (2019). Cell Form and Function: Interpreting and Controlling the Shape of Adherent Cells. *Trends in Biotechnology*, 37(4):347–357.

Qiu, X., Zhang, Y., Yang, D., Hosseinzadeh, S., Wang, L., Yuan, R., Xu, S., Ma, Y., Replogle, J., Darmanis, S., Xing, J., and Weissman, J. S. (2019). Mapping Vector Field of Single Cells. *bioRxiv*, page 696724.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P., and Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., and Human Cell Atlas Meeting Participants (2017). The Human Cell Atlas. *eLife*, 6:e27041.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.*, 44(W1):W83–9.

Reinherz, E. L. (2014). Revisiting the Discovery of the $\alpha\beta$ TCR Complex and Its Co-Receptors. *Frontiers in Immunology*, 5.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences*, 101(25):9309–9314.

Riedel, A., Shorthouse, D., Haas, L., Hall, B. A., and Shields, J. (2016). Tumor-induced stromal reprogramming drives lymph node transformation. *Nat. Immunol.*, 17(9):1118–1127.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467.

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., and Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182.

Rostom, R., Svensson, V., Teichmann, S. A., and Kar, G. (2017). Computational approaches for interpreting scRNA-seq data. *FEBS Letters*, 591(15):2213–2225.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547.

Sakaguchi, S. (2004). Naturally arising CD4+ regulatory t cells for immunologic self-tolerance and negative control of immune responses. *Annu. Rev. Immunol.*, 22:531–562.

Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M., and Toda, M. (1995). Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *The Journal of Immunology*, 155(3):1151–1164.

Sasagawa, Y., Danno, H., Takada, H., Ebisawa, M., Tanaka, K., Hayashi, T., Kurisaki, A., and Nikaido, I. (2018). Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biology*, 19(1):29.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, 33(5):495–502.

Schiering, C., Krausgruber, T., Chomka, A., Fröhlich, A., Adelmann, K., Wohlfert, E. A., Pott, J., Griseri, T., Bollrath, J., Hegazy, A. N., Harrison, O. J., Owens, B. M. J., Löhning, M., Belkaid, Y., Fallon, P. G., and Powrie, F. (2014). The alarmin IL-33 promotes regulatory t-cell function in the intestine. *Nature*, 513(7519):564–568.

Schmitt, N. and Ueno, H. (2015). Regulation of human helper T cell subset differentiation by cytokines. *Current Opinion in Immunology*, 34:130–136.

Schwann, T. (1847). *Microscopical researches into the accordance in the structure and growth of animals and plants*. Sydenham Society.

Scialdone, A., Natarajan, K. N., Saraiva, L. R., Proserpio, V., Teichmann, S. A., Stegle, O., Marioni, J. C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61.

Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N. K., Macaulay, I. C., Marioni, J. C., and Göttgens, B. (2016). Resolving early mesoderm diversification through single-cell expression profiling. *Nature*, 535(7611):289–293.

Scott, C. L., T'Jonck, W., Martens, L., Todorov, H., Sichien, D., Soen, B., Bonnardel, J., Prijck, S. D., Vandamme, N., Cannoodt, R., Saelens, W., Vanneste, B., Toussaint, W., Bleser, P. D., Takahashi, N., Vandenabeele, P., Henri, S., Pridans, C., Hume, D. A., Lambrecht, B. N., Baetselier, P. D., Milling, S. W. F., Ginderachter, J. A. V., Malissen,

B., Berx, G., Beschin, A., Saeys, Y., and Guilliams, M. (2018). The Transcription Factor ZEB2 Is Required to Maintain the Tissue-Specific Identities of Macrophages. *Immunity*, 49(2):312–325.e5.

Sefik, E., Geva-Zatorsky, N., Oh, S., Konnikova, L., Zemmour, D., McGuire, A. M., Burzyn, D., Ortiz-Lopez, A., Lobera, M., Yang, J., Ghosh, S., Earl, A., Snapper, S. B., Jupp, R., Kasper, D., Mathis, D., and Benoist, C. (2015). MUCOSAL IMMUNOLOGY. individual intestinal symbionts induce a distinct population of rorγ^+ regulatory T cells. *Science*, 349(6251):993–997.

Segal, J. M., Kent, D., Wesche, D. J., Ng, S. S., Serra, M., Oulès, B., Kar, G., Emerton, G., Blackford, S. J. I., Darmanis, S., Miquel, R., Luong, T. V., Yamamoto, R., Bonham, A., Jassem, W., Heaton, N., Vigilante, A., King, A., Sancho, R., Teichmann, S., Quake, S. R., Nakauchi, H., and Rashid, S. T. (2019). Single cell analysis of human foetal liver captures the transcriptional profile of hepatobiliary hybrid progenitors. *Nature Communications*, 10(1):1–14.

Segerstolpe, A., Palasantza, A., Eliasson, P., Andersson, E.-M., Andreasson, A.-C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M. K., Smith, D. M., Kasper, M., Ämmälä, C., and Sandberg, R. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, 24(4):593–607.

Sharma, A. and Rudra, D. (2018). Emerging Functions of Regulatory T Cells in Tissue Homeostasis. *Frontiers in Immunology*, 9.

Sharma, M. D., Huang, L., Choi, J.-H., Lee, E.-J., Wilson, J. M., Lemos, H., Pan, F., Blazar, B. R., Pardoll, D. M., Mellor, A. L., Shi, H., and Munn, D. H. (2013). An inherently bifunctional subset of foxp3+ T helper cells is controlled by the transcription factor eos. *Immunity*, 38(5):998–1012.

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., Adiconis, X., Levin, J. Z., Nemesh, J., Goldman, M., McCarroll, S. A., Cepko, C. L., Regev, A., and Sanes, J. R. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, 166(5):1308–1323.e30.

Sherwood, A. M., Emerson, R. O., Scherer, D., Habermann, N., Buck, K., Staffa, J., Desmarais, C., Halama, N., Jaeger, D., Schirmacher, P., Herpel, E., Kloor, M., Ulrich, A., Schneider, M., Ulrich, C. M., and Robins, H. (2013). Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue. *Cancer Immunol. Immunother.*, 62(9):1453–1461.

Shin, D., Lee, W., Lee, J. H., and Bang, D. (2019). Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. *Science Advances*, 5(5):eaav2249.

Shulse, C. N., Cole, B. J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M., Turco, G. M., Zhu, Y., O'Malley, R. C., Brady, S. M., and Dickel, D. E. (2019). High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types. *Cell Reports*, 27(7):2241–2247.e4.

Shyamsundar, R., Kim, Y. H., Higgins, J. P., Montgomery, K., Jorden, M., Sethuraman, A., van de Rijn, M., Botstein, D., Brown, P. O., and Pollack, J. R. (2005). A DNA microarray survey of gene expression in normal human tissues. *Genome Biology*, 6(3):R22.

Simoes, A. E., Di Lorenzo, B., and Silva-Santos, B. (2018). Molecular Determinants of Target Cell Recognition by Human $\gamma\delta$ T Cells. *Frontiers in Immunology*, 9.

Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., Herbst, R. H., Rogel, N., Slyper, M., Waldman, J., Sud, M., Andrews, E., Velonias, G., Haber, A. L., Jagadeesh, K., Vickovic, S., Yao, J., Stevens, C., Dionne, D., Nguyen, L. T., Villani, A.-C., Hofree, M., Creasey, E. A., Huang, H., Rozenblatt-Rosen, O., Garber, J. J., Khalili, H., Desch, A. N., Daly, M. J., Ananthakrishnan, A. N., Shalek, A. K., Xavier, R. J., and Regev, A. (2019). Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell*, 178(3):714–730.e22.

Sohni, A., Tan, K., Song, H.-W., Burow, D., de Rooij, D. G., Laurent, L., Hsieh, T.-C., Rabah, R., Hammoud, S. S., Vicini, E., and Wilkinson, M. F. (2019). The Neonatal and Adult Human Testis Defined at the Single-Cell Level. *Cell Reports*, 26(6):1501–1517.e4.

Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., DeMeo, D. L., Quackenbush, J., Glass, K., and Kuijjer, M. L. (2017). Understanding Tissue-Specific Gene Regulation. *Cell Reports*, 21(4):1077–1088.

Soneson, C. and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261.

Stein-O'Brien, G. L., Clark, B. S., Sherman, T., Zibetti, C., Hu, Q., Sealfon, R., Liu, S., Qian, J., Colantuoni, C., Blackshaw, S., Goff, L. A., and Fertig, E. J. (2019). Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Systems*, 8(5):395–411.e8.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, W. M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology*, 19(1):224.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21.

Stubbington, M. J. T., Lönnberg, T., Proserpio, V., Clare, S., Speak, A. O., Dougan, G., and Teichmann, S. A. (2016). T cell fate and clonality inference from single-cell transcriptomes. *Nature Methods*, 13(4):329–332.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Svensson, V. and Beltrame, E. d. V. (2019). A curated database reveals trends in single cell transcriptomics. *bioRxiv*, page 742304.

Svensson, V., Beltrame, E. d. V., and Pachter, L. (2019). Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. *bioRxiv*, page 762773.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604.

Takeda, A., Hollmén, M., Dermadi, D., Pan, J., Brulois, K. F., Kaukonen, R., Lönnberg, T., Boström, P., Koskivuo, I., Irjala, H., Miyasaka, M., Salmi, M., Butcher, E. C., and Jalkanen, S. (2019). Single-Cell Survey of Human Lymphatics Unveils Marked Endothelial Cell Heterogeneity and Mechanisms of Homing for Neutrophils. *Immunity*.

Tan, Y. and Cahan, P. (2018). SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. *bioRxiv*, page 508085.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382.

Taylor, D. M., Aronow, B. J., Tan, K., Bernt, K., Salomonis, N., Greene, C. S., Frolova, A., Henrickson, S. E., Wells, A., Pei, L., Jaiswal, J. K., Whitsett, J., Hamilton, K. E., MacParland, S. A., Kelsen, J., Heuckeroth, R. O., Potter, S. S., Vella, L. A., Terry, N. A., Ghanem, L. R., Kennedy, B. C., Helbig, I., Sullivan, K. E., Castelo-Soccio, L., Kreigstein, A., Herse, F., Nawijn, M. C., Koppelman, G. H., Haendel, M., Harris, N. L., Rokita, J. L., Zhang, Y., Regev, A., Rozenblatt-Rosen, O., Rood, J. E., Tickle, T. L., Vento-Tormo, R., Alimohamed, S., Lek, M., Mar, J. C., Loomes, K. M., Barrett, D. M., Uapinyoying, P., Beggs, A. H., Agrawal, P. B., Chen, Y.-W., Muir, A. B., Garmire, L. X., Snapper, S. B., Nazarian, J., Seeholzer, S. H., Fazelinia, H., Singh, L. N., Faryabi, R. B., Raman, P., Dawany, N., Xie, H. M., Devkota, B., Diskin, S. J., Anderson, S. A., Rappaport, E. F., Peranteau, W., Wikenheiser-Brokamp, K. A., Teichmann, S., Wallace, D., Peng, T., Ding, Y.-y., Kim, M. S., Xing, Y., Kong, S. W., Bönnemann, C. G., Mandl, K. D., and White, P. S. (2019). The Pediatric Cell Atlas: Defining the Growth Phase of Human Development at Single-Cell Resolution. *Developmental Cell*, 49(1):10–29.

The Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45:1113–1120.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Thome, J. J. C., Bickham, K. L., Ohmura, Y., Kubota, M., Matsuoka, N., Gordon, C., Granot, T., Griesemer, A., Lerner, H., Kato, T., and Farber, D. L. (2015). Early-life compartmentalization of human T cell differentiation and regulatory function in mucosal and lymphoid tissues. *Nat. Med.*, 22(1):72–77.

Titsias, M. K. and Lawrence, N. D. (2010). Bayesian gaussian process latent variable model. *Artif. Intell.*, 9:844–851.

Traag, V. A., Waltman, L., and Eck, N. J. v. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386.

Trzpis, M., McLaughlin, P. M., de Leij, L. M., and Harmsen, M. C. (2007). Epithelial Cell Adhesion Molecule. *The American Journal of Pathology*, 171(2):386–395.

Uhlig, H. H., Coombes, J., Mottet, C., Izcue, A., Thompson, C., Fanger, A., Tannapfel, A., Fontenot, J. D., Ramsdell, F., and Powrie, F. (2006). Characterization of Foxp3+CD4+CD25+ and IL-10-secreting CD4+CD25+ T cells during cure of colitis. *J. Immunol.*, 177(9):5852–5860.

van den Brink, S. C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C. S., Robin, C., and van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods*, 14(10):935–936.

Various (2017). What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? *Cell Systems*, 4(3):255–259.

Various (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367.

Vasanthakumar, A., Liao, Y., Teh, P., Pascutti, M. F., Oja, A. E., Garnham, A. L., Gloury, R., Tempany, J. C., Sidwell, T., Cuadrado, E., Tuijnenburg, P., Kuijpers, T. W., Lalaoui, N., Mielke, L. A., Bryant, V. L., Hodgkin, P. D., Silke, J., Smyth, G. K., Nolte, M. A., Shi, W., and Kallies, A. (2017). The TNF receptor Superfamily-NF-κB axis is critical to maintain effector regulatory T cells in lymphoid and non-lymphoid tissues. *Cell Rep.*, 20(12):2906–2920.

Vasanthakumar, A., Moro, K., Xin, A., Liao, Y., Gloury, R., Kawamoto, S., Fagarasan, S., Mielke, L. A., Afshar-Sterle, S., Masters, S. L., Nakae, S., Saito, H., Wentworth, J. M., Li, P., Liao, W., Leonard, W. J., Smyth, G. K., Shi, W., Nutt, S. L., Koyasu, S., and Kallies, A. (2015). The transcriptional regulators IRF4, BATF and IL-33 orchestrate development and maintenance of adipose tissue–resident regulatory T cells. *Nat. Immunol.*, 16(3):276–285.

Veiga-Fernandes, H. and Mucida, D. (2016). Neuro-Immune Interactions at Barrier Surfaces. *Cell*, 165(4):801–811.

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J.-E., Stephenson, E., Polański, K., Goncalves, A., Gardner, L., Holmqvist, S., Henriksson, J., Zou, A., Sharkey, A. M., Millar, B., Innes, B., Wood, L., Wilbrey-Clark, A., Payne, R. P., Ivarsson, M. A., Lisgo, S., Filby, A., Rowitch, D. H., Bulmer, J. N., Wright, G. J., Stubbington, M. J. T., Haniffa, M., Moffett, A., and Teichmann, S. A. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, 563(7731):347.

Verboom, K., Everaert, C., Bolduc, N., Livak, K. J., Yigit, N., Rombaut, D., Anckaert, J., Lee, S., Venø, M. T., Kjems, J., Speleman, F., Mestdagh, P., and Vandesompele, J. (2019). SMARTer single cell total RNA sequencing. *Nucleic Acids Research*, 47(16):e93–e93.

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J. F., Gould, J., Ronaghi, M., Frisén, J., Lundeberg, J., Regev, A., and Ståhl, P. L. (2019). High-density spatial transcriptomics arrays for in situ tissue profiling. *bioRxiv*, page 563338.

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., Jager, P. L. D., Rozenblatt-Rosen, O., Lane, A. A., Haniffa, M., Regev, A., and Hacohen, N. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335):eaah4573.

Wagner, F. and Yanai, I. (2018). Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. *bioRxiv*, page 456129.

Wang, C.-C., Jamal, L., and Janes, K. A. (2012). Normal morphogenesis of epithelial tissues and progression of epithelial tumors. *Wiley interdisciplinary reviews. Systems biology and medicine*, 4(1):51–78.

Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Conley, V. B., MacMullan, H., and Zhang, N. R. (2018). Transfer learning in single-cell transcriptomics improves data denoising and pattern discovery. *bioRxiv*, page 457879.

Wang, Y., Hoinka, J., and Przytycka, T. M. (2019). Subpopulation Detection and Their Comparative Analysis across Single-Cell Experiments with scPopCorn. *Cell Systems*, 8(6):506–513.e5.

Wang, Y. J., Schug, J., Won, K.-J., Liu, C., Naji, A., Avrahami, D., Golson, M. L., and Kaestner, K. H. (2016). Single-Cell Transcriptomics of the Human Endocrine Pancreas. *Diabetes*, 65(10):3028–3038.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Watcham, S., Kucinski, I., and Gottgens, B. (2019). New Insights into Haematopoietic Differentiation Landscapes from scRNA-seq. *Blood*, pages blood–2018–08–835355.

Weaver, C. T., Elson, C. O., Fouser, L. A., and Kolls, J. K. (2013). The Th17 Pathway and Inflammatory Diseases of the Intestines, Lungs, and Skin. *Annual Review of Pathology: Mechanisms of Disease*, 8(1):477–512.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17.

Wheaton, J. D. and Ciofani, M. (2019). JunB controls intestinal effector programs in regulatory T cells. *bioRxiv*, page 772194.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15.

Wong, M. T., Ong, D. E. H., Lim, F. S. H., Teng, K. W. W., McGovern, N., Narayanan, S., Ho, W. Q., Cerny, D., Tan, H. K. K., Anicete, R., Tan, B. K., Lim, T. K. H., Chan, C. Y., Cheow, P. C., Lee, S. Y., Takano, A., Tan, E.-H., Tam, J. K. C., Tan, E. Y., Chan, J. K. Y., Fink, K., Bertoletti, A., Ginhoux, F., Curotto de Lafaille, M. A., and Newell, E. W. (2016a). A High-Dimensional Atlas of Human T Cell Diversity Reveals Tissue-Specific Trafficking and Cytokine Signatures. *Immunity*, 45(2):442–456.

Wong, S. C., Gatt, A., Stamatescu, V., and McDonnell, M. D. (2016b). Understanding data augmentation for classification: when to warp? *arXiv:1609.08764 [cs]*. arXiv: 1609.08764.

Wu, Y. E., Pan, L., Zuo, Y., Li, X., and Hong, W. (2017). Detecting activated cell populations using Single-Cell RNA-Seq. *Neuron*, 96(2):313–329.e6.

Xie, P., Gao, M., Wang, C., Zhang, J., Noel, P., Yang, C., Von Hoff, D., Han, H., Zhang, M. Q., and Lin, W. (2019). SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. *Nucleic Acids Research*, 47(8):e48–e48.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., and Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–659.

Yin, Y., Wang, X. X., and Mariuzza, R. A. (2012). Crystal structure of a complete ternary complex of T-cell receptor, peptide–MHC, and CD4. *Proceedings of the National Academy of Sciences*, 109(14):5405–5410.

Young, M. D., Mitchell, T. J., Braga, F. A. V., Tran, M. G. B., Stewart, B. J., Ferdinand, J. R., Collord, G., Botting, R. A., Popescu, D.-M., Loudon, K. W., Vento-Tormo, R., Stephenson, E., Cagan, A., Farndon, S. J., Velasco-Herrera, M. D. C., Guzzo, C., Richoz, N., Mamanova, L., Aho, T., Armitage, J. N., Riddick, A. C. P., Mushtaq, I., Farrell, S., Rampling, D., Nicholson, J., Filby, A., Burge, J., Lisgo, S., Maxwell, P. H., Lindsay, S., Warren, A. Y., Stewart, G. D., Sebire, N., Coleman, N., Haniffa, M., Teichmann, S. A., Clatworthy, M., and Behjati, S. (2018). Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science*, 361(6402):594–599.

Yu, Y., Ma, X., Gong, R., Zhu, J., Wei, L., and Yao, J. (2018). Recent advances in CD8+ regulatory T cell research (Review). *Oncology Letters*, 15(6):8187–8194.

Zemmour, D., Zilionis, R., Kiner, E., Klein, A. M., Mathis, D., and Benoist, C. (2018). Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nature Immunology*, 19(3):291–301.

Zhang, A. W., O'Flanagan, C., Chavez, E., Lim, J. L., McPherson, A., Wiens, M., Walters, P., Chan, T., Hewitson, B., Lai, D., Mottok, A., Sarkozy, C., Chong, L., Aoki, T., Wang, X., Weng, A. P., McAlpine, J. N., Aparicio, S., Steidl, C., Campbell, K. R., and Shah, S. P. (2019a). Probabilistic cell type assignment of single-cell transcriptomic data reveals spatiotemporal microenvironment dynamics in human cancers. *bioRxiv*, page 521914.

Zhang, H., Kong, H., Zeng, X., Guo, L., Sun, X., and He, S. (2014). Subsets of regulatory T cells and their roles in allergy. *Journal of Translational Medicine*, 12(1):125.

Zhang, L., Yu, X., Zheng, L., Zhang, Y., Li, Y., Fang, Q., Gao, R., Kang, B., Zhang, Q., Huang, J. Y., Konno, H., Guo, X., Ye, Y., Gao, S., Wang, S., Hu, X., Ren, X., Shen, Z., Ouyang, W., and Zhang, Z. (2018). Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature*, 564(7735):268–272.

Zhang, Z., Luo, D., Zhong, X., Choi, J. H., Ma, Y., Mahrt, E., Guo, W., Stawiski, E. W., Wang, S., Modrusan, Z., Seshagiri, S., Kapur, P., Wang, X., Hon, G. C., Brugarolas, J., and Wang, T. (2019b). SCINA: Semi-Supervised Analysis of Single Cells in silico. *bioRxiv*, page 559872.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049.

Zimmermann, B., Robert, N. S. M., Technau, U., and Simakov, O. (2019). Ancient animal genome architecture reflects cell type identities. *Nature Ecology & Evolution*, 3(9):1289–1293.

Appendix A

Additional information to Chapter 2

This Appendix contains supplementary information for Chapter 2, including experimental methods and supplementary figures.

A.1 Additional Experimental Methods

A.1.1 Mice

All mice were maintained under specific pathogen-free conditions at the Wellcome Genome Campus Research Support Facility (Cambridge, UK) and at the Kennedy Institute for Rheumatology (Oxford, UK). All procedures were in accordance with the Animals Scientific Procedures Act 1986. For steady-state experiments, the Foxp3-GFP-KI mouse reporter line (Bettelli et al., 2006) was used. The melanoma challenge was performed in Foxp3-IRES-GFP genetically targeted reporter mice (Haribhai et al., 2007) purchased from The Jackson Laboratory (stock no. 006772). In both cases, 6-14 week-old females were used.

A.1.2 Human samples

Human skin and blood samples were obtained from patients undergoing breast reduction plastic surgeries (REC approval number: 08/H0906/95+5). Surgical-resection specimens were obtained from patients attending the John Radcliffe Hospital Gastroenterology Unit (Oxford, UK). These specimens were obtained from normal regions of bowel adjacent to resected colorectal tumours from patients undergoing surgery. Informed, written consent was obtained from all donors. Human experimental protocols

were approved by the NHS Research Ethics System (Reference number:11/YH/0020). Further details concerning patients and tumours can be found in Table A.6.

A.1.3 Murine leukocytes isolation in steady-state skin dataset

To isolate leukocytes from ear tissue, ears were removed at the base, split into halves and cut into very small pieces. Tissue was digested in 3.5ml RPMI medium (GIBCO) with 0.1% BSA, 15mM Hepes, 1mg/ml collagenase D (Roche) and 450μg/ml Liberase TL (Roche) for 60 minutes at 37°C in a shaking incubator at 200rpm. Digested tissue was passed through an 18G needle to further disrupt the tissue and release cells. Cells were filtered through a 70μm cell strainer, and the digestion was terminated by addition of ice-cold RPMI containing 0.1% BSA (Sigma-Aldrich) and 5mM EDTA (Invitrogen). A three-layer (30, 40, 70%) Percoll (GE Healthcare) density-gradient was used to enrich for the lymphocytes. Cells obtained from the digestion were layered in the 30% layer on top of the 40% and 70% layers, and centrifuged for 20 minutes at 1800rpm without brake. Cells at the 40/70% interface were collected for the subsequent analysis. Cell suspensions from spleen and bLN were prepared as described previously (Uhlig et al., 2006).

A.1.4 Murine leukocytes isolation in steady-state colon dataset

Colons were washed twice in RPMI medium (GIBCO) with 0.1% BSA (Sigma-Aldrich) and 5mM EDTA (Invitrogen) in a shaking incubator at 200rpm at 37°C to remove epithelial cells. The tissue was then digested for an hour in RPMI with 10% FCS, 15mM Hepes (GIBCO) and 100U/ml collagenase VIII (Sigma-Aldrich). Digestion was terminated by addition of ice-cold RPMI with 10% FCS (Sigma-Aldrich) and 5mM EDTA (Invitrogen). Leukocyte enrichment and suspension was obtained as described in the previous paragraph.

A.1.5 Melanoma induction and cell isolation

The melanoma induction experiments were performed in accordance with UK Home Office regulations under Project License PPL 80/2574. The protocol used was adapted from a previous publication (Riedel et al., 2016). For syngeneic tumours, 2.5 $\times$ 105 B16.F10 melanoma cells (ATCC) were inoculated subcutaneously into the shoulder region of 6- to 14-week-old female Foxp3-IRES-GFP mice (Haribhai et al., 2007). Animals were excluded if tumours failed to form or if health concerns were reported.

Control Foxp3-IRES-GFP mice were injected with 50 μl PBS. Animals were culled after 11 days. Tumours, tumour-draining (brachial) lymph nodes and spleen were isolated for subsequent analysis. PBS-injected and steady-state skin, draining lymph nodes (bLN) and spleen were collected from control mice. Tumour and PBS-injected skin were mechanically disrupted and digested in a 1ml mixture of 1 mg/ml collagenase A (Roche) and 0.4 mg/ml DNase I (Roche) in PBS (solution A) at 37°C for 1h with 600rpm rotation. 1ml of PBS containing 1mg/ml Collagenase D (Roche) and 0.4 mg/ml DNase I (Roche) (solution B) was then added to each sample, which returned to 37 °C for 1h with 600 rpm rotation. Lymph nodes were digested for 30min in 500μl of solution A, and for further 30min after the addition of 500μl of solution B. EDTA (Invitrogen) at the final concentration of 10mM was added to all samples. Spleens were processed as described previously (Uhlig et al., 2006). Suspensions were passed through a 70 μm cell strainer before immunostaining. Samples from different animals were kept separated throughout processing and sorting.

A.1.6 Isolation of human CD4+ T cells

Isolation of leukocytes from human skin

Plastic surgery skin included reticular dermis to the depth of the fat layer. The upper 200 microns of skin were harvested using a split skin graft knife. Whole skin was digested in RPMI 1640 with 100IU/ml penicillin, 100μ/ml streptomycin, 2mM L-glutamine (GIBCO), 10% FCS (Sigma-Aldrich) and 1.6mg/ml type IV collagenase (Worthington-Biochemical) for 12-16 hours at 37°C and 5% CO2. Digest was passed repeatedly through a 10ml pipette until no visible material remained. To yield a single-cell suspension, digest was passed through a 100-micron filter into a polypropylene sorting tube.

Isolation of leukocytes from human colon

Normal regions of bowel adjacent to resected colorectal tumours were prepared as previously described, with minor modifications (Bettelli et al., 2006; Geremia et al., 2011). In brief, mucosa was dissected and washed in 1 mM dithiothreitol (DTT) (Sigma-Aldrich) solution for 15 min at room temperature to remove mucus. Specimens were then washed three times in 0.75 mM EDTA (Invitrogen) to deplete epithelial crypts and were digested for 2h in 0.1 mg/ml collagenase A solution (Roche). For enrichment of mononuclear cells, digests were centrifuged for 30 min at

500g in a four-layer Percoll (GE Healthcare) gradient and collected at the 40%/60% interface.

Peripheral blood mononuclear cell isolation

10mL blood from skin donors were collected into EDTA (Invitrogen). Density centrifugation with Lymphoprep (STEMCELL Technologies) was performed according to manufacturer's instructions. Recovered cells were cryopreserved by pelleting and resuspending in 1ml heat-inactivated fetal calf serum containing 10% DMSO, and storing at -80ºC. Cryovials were later thawed in water bath, then rapidly being transferred to warmed medium (RPMI 1640 (GIBCO) with 100IU/ml penicillin, 100μg/ml streptomycin, 2mM L-glutamine (GIBCO), 10% FCS (Sigma-Aldrich)) and filtered through a 100-μm filter.

A.1.7 Flow cytometry and single-cell RNA sequencing

Mouse and human cell suspensions were sorted as described in Figure 2.1A, Figure 2.4A, Figure 2.5A, and Figure A.1A.

Droplet-based scRNA-seq datasets were produced using a Chromium system (10x Genomics), referred to as 10x. Cell populations of interest were sorted, manually counted, and their concentrations adjusted to enable the capture of $\tilde{5}000$ cells (except for skin Treg and Tmem cells, for which we aimed to capture $\tilde{3}00$ each). The standard protocol for the 10x single cell 3' kit (V2 chemistry) was followed and each cell population loaded onto a separate chip inlet. We ran each sample on one lane of Illumina HiSeq 4000, following manufacturer's recommendations.

Two plate-based scRNA-seq datasets: the "colon dataset", including Treg and Tmem cells from colon, mLN and spleen, and the "skin dataset" from skin, bLN and spleen. Single cells were sorted in 2μl of Lysis Buffer (1:20 solution of RNase Inhibitor (Clontech) in 0.2% v/v Triton X-100 (Sigma-Aldrich)) in 96 well plates, spun down and immediately frozen at -80ºC. Smart-seq2 protocol (Picelli et al., 2014) was largely followed to obtain mRNA libraries from single cells. Oligo-dT primer, dNTPs (ThermoFisher) and ERCC RNA Spike-In Mix (1:50,000,000 final dilution, Ambion) were then added. Reverse Transcription and PCR were performed as previously published (Picelli et al., 2014), using 50U of SMARTScribe™ Reverse Transcriptase (Clontech). The cDNA libraries for sequencing were prepared using Nextera XT DNA Sample Preparation Kit (Illumina), according to the protocol supplied by Fluidigm. Libraries from single cells were pooled and purified using AMPure XP beads (Beckman

Coulter). Pooled samples were sequenced on an Illumina HiSeq 2500 (paired-end 100-bp reads) or Illumina HiSeq 2000 v4 chemistry (paired-end 75-bp reads) aiming at an average depth of 1 million reads/cell.

A.2 Supplementary Tables and Figures

Table A.1: Batch details for the Mouse steady-state Smart-seq2 data.

Experiment	Tissue.Cell Type	Condition	Donor	Plate/Chip	Plate/Chip date	Library Date
mouse_colon	spleen.Treg	Steady-state	Pool	mouse_colon_5	25/05/2015	17/09/2015
mouse_colon	colon.Treg	Steady-state	Pool	mouse_colon_14	27/05/2015	17/09/2015
mouse_colon	colon.Tmem	Steady-state	Pool	mouse_colon_18	28/05/2015	17/09/2015
mouse_colon	LN.Treg	Steady-state	Pool	mouse_colon_10	25/05/2015	17/09/2015
mouse_colon	LN.Tmem	Steady-state	Pool	mouse_colon_11	28/05/2015	17/09/2015
mouse_colon	LN.Treg	Steady-state	Pool	mouse_colon_9	21/09/2015	10/06/2015
mouse_colon	colon.Treg	Steady-state	Pool	mouse_colon_15	29/09/2015	10/06/2015
mouse_colon	colon.Treg	Steady-state	Pool	mouse_colon_16	30/09/2015	10/06/2015
mouse_colon	spleen.Tmem	Steady-state	Pool	mouse_colon_4	30/09/2015	10/06/2015
mouse_colon	spleen.Treg	Steady-state	Pool	mouse_colon_7	29/09/2015	10/06/2015
mouse_colon	spleen.Treg	Steady-state	Pool	mouse_colon_7	29/09/2015	10/01/2015
mouse_colon	LN.Tmem	Steady-state	Pool	mouse_colon_11	28/05/2015	10/06/2015
mouse_colon	colon.Tmem	Steady-state	Pool	mouse_colon_18	28/05/2015	10/06/2015
mouse_colon	colon.Treg	Steady-state	Pool	mouse_colon_16	30/09/2015	10/01/2015
mouse_colon	LN.Treg	Steady-state	Pool	mouse_colon_9	21/09/2015	10/01/2015
mouse_colon	colon.Treg	Steady-state	Pool	mouse_colon_15	29/09/2015	10/01/2015
mouse_colon	colon.Tmem	Steady-state	Pool	mouse_colon_18	28/05/2015	10/01/2015
mouse_colon	spleen.Tmem	Steady-state	Pool	mouse_colon_4	30/09/2015	10/01/2015
mouse_colon	LN.Tmem	Steady-state	Pool	mouse_colon_11	28/05/2015	10/01/2015
mouse_colon	spleen.Treg	Steady-state	Pool	mouse_colon_7	29/09/2015	10/08/2015
mouse_colon	LN.Treg	Steady-state	Pool	mouse_colon_9	21/09/2015	10/08/2015
mouse_colon	spleen.Tmem	Steady-state	Pool	mouse_colon_4	30/09/2015	10/08/2015
mouse_colon	LN.Tmem	Steady-state	Pool	mouse_colon_11	28/05/2015	10/08/2015
mouse_colon	colon.Treg	Steady-state	Pool	mouse_colon_15	29/09/2015	10/08/2015
mouse_colon	LN.Treg	Steady-state	Pool	mouse_colon_9	21/09/2015	10/07/2015
mouse_colon	colon.Tmem	Steady-state	Pool	mouse_colon_18	28/05/2015	10/08/2015
mouse_colon	colon.Treg	Steady-state	Pool	mouse_colon_15	29/09/2015	10/07/2015
mouse_colon	colon.Tmem	Steady-state	Pool	mouse_colon_18	28/05/2015	10/07/2015
mouse_colon	spleen.Treg	Steady-state	Pool	mouse_colon_7	29/09/2015	10/07/2015
mouse_colon	spleen.Tmem	Steady-state	Pool	mouse_colon_4	30/09/2015	10/07/2015
mouse_colon	LN.Tmem	Steady-state	Pool	mouse_colon_11	28/05/2015	10/07/2015
mouse_skin	skin.Treg	Steady-state	Pool	mouse_skin_1	08/04/2017	19/04/2017
mouse_skin	skin.Tmem	Steady-state	Pool	mouse_skin_1	08/04/2017	19/04/2017
mouse_skin	skin.Treg	Steady-state	Pool	mouse_skin_2	08/04/2017	19/04/2017
mouse_skin	skin.Tmem	Steady-state	Pool	mouse_skin_2	08/04/2017	19/04/2017
mouse_skin	skin.Treg	Steady-state	Pool	mouse_skin_3	08/04/2017	19/04/2017
mouse_skin	skin.Tmem	Steady-state	Pool	mouse_skin_3	08/04/2017	19/04/2017
mouse_skin	skin.Treg	Steady-state	Pool	mouse_skin_4	08/04/2017	18/04/2017
mouse_skin	skin.Tmem	Steady-state	Pool	mouse_skin_4	08/04/2017	18/04/2017
mouse_skin	skin.Treg	Steady-state	Pool	mouse_skin_5	08/04/2017	19/04/2017
mouse_skin	skin.Tmem	Steady-state	Pool	mouse_skin_5	08/04/2017	19/04/2017
mouse_skin	spleen.Treg	Steady-state	Pool	mouse_skin_8	08/04/2017	18/04/2017
mouse_skin	spleen.Tmem	Steady-state	Pool	mouse_skin_8	08/04/2017	18/04/2017
mouse_skin	spleen.Treg	Steady-state	Pool	mouse_skin_9	08/04/2017	18/04/2017
mouse_skin	spleen.Tmem	Steady-state	Pool	mouse_skin_9	08/04/2017	18/04/2017
mouse_skin	LN.Treg	Steady-state	Pool	mouse_skin_14	08/04/2017	19/04/2017
mouse_skin	LN.Tmem	Steady-state	Pool	mouse_skin_14	08/04/2017	19/04/2017
mouse_skin	LN.Treg	Steady-state	Pool	mouse_skin_15	08/04/2017	19/04/2017
mouse_skin	LN.Tmem	Steady-state	Pool	mouse_skin_15	08/04/2017	19/04/2017
mouse_skin	LN.Treg	Steady-state	Pool	mouse_skin_16	08/04/2017	18/04/2017
mouse_skin	LN.Tmem	Steady-state	Pool	mouse_skin_16	08/04/2017	18/04/2017

Table A.2: Batch details for the Mouse melanoma Smart-seq2 data.

Experiment	Tissue.Cell Type	Condition	Donor	Plate/Chip	Plate/Chip date	Library Date
mouse_mel	skin.Treg	Tumour	T1	mouse_mel_641	01/06/2016	NA
mouse_mel	skin.Treg	Tumour	T2	mouse_mel_641	01/06/2016	NA
mouse_mel	spleen.Treg	Control	C1	mouse_mel_637	02/06/2016	NA
mouse_mel	spleen.Treg	Control	C2	mouse_mel_637	02/06/2016	NA
mouse_mel	spleen.Treg	Control	C3	mouse_mel_637	02/06/2016	NA
mouse_mel	spleen.Treg	Control	C4	mouse_mel_644	02/06/2016	NA
mouse_mel	skin.Treg	Control	C4	mouse_mel_645	01/06/2016	NA
mouse_mel	spleen.Treg	Control	C5	mouse_mel_644	02/06/2016	NA
mouse_mel	skin.Treg	Control	C5	mouse_mel_645	01/06/2016	NA
mouse_mel	spleen.Treg	Control	C6	mouse_mel_644	02/06/2016	NA
mouse_mel	skin.Treg	Control	C6	mouse_mel_645	01/06/2016	NA
mouse_mel	skin.Treg	Tumour	T5	mouse_mel_641	01/06/2016	NA
mouse_mel	spleen.Tmem	Control	C1	mouse_mel_637	02/06/2016	NA
mouse_mel	spleen.Tmem	Control	C2	mouse_mel_637	02/06/2016	NA
mouse_mel	spleen.Tmem	Control	C3	mouse_mel_637	02/06/2016	NA
mouse_mel	spleen.Tmem	Control	C4	mouse_mel_644	02/06/2016	NA
mouse_mel	spleen.Tmem	Control	C5	mouse_mel_644	02/06/2016	NA
mouse_mel	spleen.Tmem	Control	C6	mouse_mel_644	02/06/2016	NA
mouse_mel	skin.Tmem	Control	C4	mouse_mel_645	01/06/2016	NA
mouse_mel	skin.Treg	Tumour	T1	mouse_mel_640	02/06/2016	NA
mouse_mel	spleen.Treg	Tumour	T1	mouse_mel_638	02/06/2016	NA
mouse_mel	spleen.Treg	Tumour	T2	mouse_mel_638	02/06/2016	NA
mouse_mel	skin.Treg	Tumour	T2	mouse_mel_640	02/06/2016	NA
mouse_mel	spleen.Treg	Tumour	T5	mouse_mel_638	02/06/2016	NA
mouse_mel	skin.Treg	Tumour	T5	mouse_mel_640	02/06/2016	NA
mouse_mel	skin.Treg	Control	C4	mouse_mel_648	02/06/2016	NA
mouse_mel	skin.Treg	Control	C4	mouse_mel_646	02/06/2016	NA
mouse_mel	skin.Treg	Control	C5	mouse_mel_646	02/06/2016	NA
mouse_mel	skin.Treg	Control	C6	mouse_mel_646	02/06/2016	NA
mouse_mel	spleen.Tmem	Tumour	T1	mouse_mel_638	02/06/2016	NA
mouse_mel	spleen.Tmem	Tumour	T5	mouse_mel_638	02/06/2016	NA
mouse_mel	skin.Tmem	Control	C4	mouse_mel_646	02/06/2016	NA
mouse_mel	skin.Tmem	Control	C5	mouse_mel_646	02/06/2016	NA
mouse_mel	skin.Tmem	Control	C6	mouse_mel_648	02/06/2016	NA
mouse_mel	spleen.Tmem	Tumour	T2	mouse_mel_638	02/06/2016	NA
mouse_mel	skin.Tmem	Control	C6	mouse_mel_648	01/06/2016	NA
mouse_mel	skin.Treg	Control	C6	mouse_mel_648	01/06/2016	NA
mouse_mel	LN.Treg	Tumour	T2	mouse_mel_642	02/06/2016	NA
mouse_mel	LN.Treg	Tumour	T1	mouse_mel_642	02/06/2016	NA
mouse_mel	spleen.Treg	Tumour	T1	mouse_mel_639	02/06/2016	NA
mouse_mel	spleen.Treg	Tumour	T2	mouse_mel_639	02/06/2016	NA
mouse_mel	spleen.Treg	Tumour	T5	mouse_mel_639	02/06/2016	NA
mouse_mel	LN.Treg	Tumour	T5	mouse_mel_642	02/06/2016	NA
mouse_mel	LN.Treg	Control	C4	mouse_mel_643	02/06/2016	NA
mouse_mel	LN.Treg	Control	C5	mouse_mel_643	02/06/2016	NA
mouse_mel	LN.Treg	Control	C6	mouse_mel_643	02/06/2016	NA
mouse_mel	LN.Tmem	Tumour	T1	mouse_mel_642	02/06/2016	NA
mouse_mel	LN.Tmem	Tumour	T2	mouse_mel_642	02/06/2016	NA
mouse_mel	spleen.Tmem	Tumour	T2	mouse_mel_639	02/06/2016	NA
mouse_mel	spleen.Tmem	Tumour	T5	mouse_mel_639	02/06/2016	NA
mouse_mel	LN.Tmem	Tumour	T5	mouse_mel_642	02/06/2016	NA
mouse_mel	LN.Tmem	Control	C4	mouse_mel_643	02/06/2016	NA
mouse_mel	LN.Tmem	Control	C5	mouse_mel_643	02/06/2016	NA
mouse_mel	LN.Tmem	Control	C6	mouse_mel_643	02/06/2016	NA
mouse_mel	spleen.Tmem	Tumour	T1	mouse_mel_639	02/06/2016	NA

Table A.3: Batch details for the Human steady-state Smart-seq2 data.

Experiment	Tissue.Cell Type	Condition	Donor	Plate/Chip	Plate/Chip date	Library Date
human	skin.Treg	Steady-state	skin_1	human_plate_skin_9	27/10/2015	11/11/2015
human	skin.Tem	Steady-state	skin_1	human_plate_skin_8	20/10/2015	11/11/2015
human	skin.Tcm	Steady-state	skin_1	human_plate_skin_7	28/10/2015	11/11/2015
human	blood.Treg	Steady-state	skin_1	human_plate_skin_3	23/09/2015	11/11/2015
human	blood.Tem	Steady-state	skin_1	human_plate_skin_2	28/10/2015	11/11/2015
human	blood.Tcm	Steady-state	skin_1	human_plate_skin_1	27/10/2015	11/11/2015
human	skin.Tcm	Steady-state	skin_2	human_743	09/06/2016	date_lib_skin_2
human	skin.Treg	Steady-state	skin_2	human_741	09/06/2016	date_lib_skin_2
human	skin.Tem	Steady-state	skin_2	human_743	09/06/2016	date_lib_skin_2
human	blood.Treg	Steady-state	skin_2	human_741	09/06/2016	date_lib_skin_2
human	skin.Treg	Steady-state	skin_3	human_745	10/06/2016	date_lib_skin_2
human	skin.Tcm	Steady-state	skin_3	human_747	10/06/2016	date_lib_skin_2
human	skin.Tem	Steady-state	skin_3	human_747	10/06/2016	date_lib_skin_2
human	blood.Treg	Steady-state	skin_3	human_745	10/06/2016	date_lib_skin_2
human	blood.Tcm	Steady-state	skin_3	human_747	10/06/2016	date_lib_skin_2
human	blood.Tem	Steady-state	skin_3	human_747	10/06/2016	date_lib_skin_2
human	blood.Tcm	Steady-state	skin_2	human_743	09/06/2016	date_lib_skin_2
human	blood.Tem	Steady-state	skin_2	human_743	09/06/2016	date_lib_skin_2
human	colon.Treg	Steady-state	colon_1	human_2	15/11/2016	07/12/2016
human	colon.Treg	Steady-state	colon_1	human_1	17/09/2016	07/12/2016
human	colon.Tcm	Steady-state	colon_1	human_5	15/11/2016	07/12/2016
human	colon.Treg	Steady-state	colon_2	human_1	15/11/2016	07/12/2016
human	colon.Tem	Steady-state	colon_2	human_7	15/11/2016	07/12/2016
human	colon.Tem	Steady-state	colon_1	human_7	15/11/2016	07/12/2016
human	colon.Tcm	Steady-state	colon_2	human_4	15/11/2016	07/12/2016
human	colon.Treg	Steady-state	colon_1	human_2	15/11/2016	01/12/2016
human	colon.Treg	Steady-state	colon_1	human_1	17/09/2016	01/12/2016
human	colon.Tcm	Steady-state	colon_2	human_4	15/11/2016	01/12/2016
human	colon.Tem	Steady-state	colon_2	human_7	15/11/2016	01/12/2016
human	colon.Treg	Steady-state	colon_2	human_2	17/09/2016	07/12/2016
human	colon.Treg	Steady-state	colon_2	human_1	15/11/2016	01/12/2016
human	skin.Treg	Steady-state	skin_2	human_742	NA	date_lib_skin_2
human	skin.Tem	Steady-state	skin_2	human_744	NA	date_lib_skin_2
human	skin.Tcm	Steady-state	skin_2	human_744	NA	date_lib_skin_2
human	blood.Treg	Steady-state	skin_2	human_742	NA	date_lib_skin_2
human	skin.Treg	Steady-state	skin_3	human_746	NA	date_lib_skin_2
human	skin.Tcm	Steady-state	skin_3	human_748	NA	date_lib_skin_2
human	skin.Tem	Steady-state	skin_3	human_748	NA	date_lib_skin_2
human	blood.Treg	Steady-state	skin_3	human_746	NA	date_lib_skin_2
human	blood.Tcm	Steady-state	skin_3	human_748	NA	date_lib_skin_2
human	blood.Tem	Steady-state	skin_3	human_748	NA	date_lib_skin_2
human	blood.Tem	Steady-state	skin_2	human_744	NA	date_lib_skin_2
human	blood.Tcm	Steady-state	skin_2	human_744	NA	date_lib_skin_2

Table A.4: Batch details for the Mouse steady-state Chromium 10x data.

Experiment	Tissue.Cell Type	Condition	Donor	Plate/Chip	Plate/Chip date	Library Date
mouse_10x	skin.Treg	Steady-state	Pool	chip 2	date_10x_run	date_10x_run
mouse_10x	skin.Tmem	Steady-state	Pool	chip 2	date_10x_run	date_10x_run
mouse_10x	colon.Treg	Steady-state	Pool	chip 1	date_10x_run	date_10x_run
mouse_10x	colon.Tmem	Steady-state	Pool	chip 1	date_10x_run	date_10x_run
mouse_10x	mLN.Treg	Steady-state	Pool	chip 1	date_10x_run	date_10x_run
mouse_10x	mLN.Tmem	Steady-state	Pool	chip 1	date_10x_run	date_10x_run
mouse_10x	bLN.Treg	Steady-state	Pool	chip 1	date_10x_run	date_10x_run
mouse_10x	bLN.Tmem	Steady-state	Pool	chip 1	date_10x_run	date_10x_run
mouse_10x	spleen.Treg	Steady-state	Pool	chip 1	date_10x_run	date_10x_run
mouse_10x	spleen.Tmem	Steady-state	Pool	chip 1	date_10x_run	date_10x_run

Table A.5: Quality control criteria for filtering single cell transcriptomes in each dataset, parameters for dimensionality reduction and QC rejection fractions. Cells were kept if they passed all these filters (see Methods). Related to Figure 2.1

	Mouse Colon	Mouse Skin	Mouse Melanoma	Human Skin/Colon	Mouse 10x
Protocol	Smart-seq2	Smart-seq2	Smart-seq2	Smart-seq2	Chromium (10x)
Maximum mitochondrial reads (%)	10	10	10	20	Not Used
Maximum ERCC-derived reads (%)	25	25	25	50	Not Used
Maximum unmapped reads (%)	30	30	30	60	Not Used
Minimum number of detected genes	1750	1750	1750	1000	700
Minimum number of mapped reads/UMI	250000	250000	250000	100000	1000
Contains TCR reads (TraCeR)	Y	Y	Y	Y	Not Used
Number of PCs for tSNE/clustering	20	20	20	20	30
tSNE perplecity	30	30	30	30	30
QC rejection fraction	0.23	0.11	0.33	0.15	0.01
TCR rejection fraction (after QC)	0.16	0.11	0.20	0.28	Not Used
Maximum number of detected genes	Not Used	Not Used	Not Used	Not Used	3500
Maximum number of UMI	Not Used	Not Used	Not Used	Not Used	15000
Clustering rejection fraction (after QC)	Not Used	Not Used	Not Used	Not Used	0.09

Table A.6: Information on human donors with biological material included in this study. Related to Figure 2.5

	skin_1	skin_2	skin_3	colon_1	colon_2
Tissue	Skin	Skin	Skin	Colon	Colon
Age	-	-	-	64	62
Sex	F	F	F	F	M
Pathology and location	Breast reduction; Breast	Breast reduction; Breast	Breast reduction; Breast	adenocarcinoma; Caecum	Tubilovillous adenoma; rectum
Tumour stage	-	-	-	PT3 N0(0/23) M0 L0 V0 R0 Duke's B	PT0
Date of diagnosis	-	-	-	-	Oct/2015
Observations	Matching blood sample	Matching blood sample	Matching blood sample	-	-

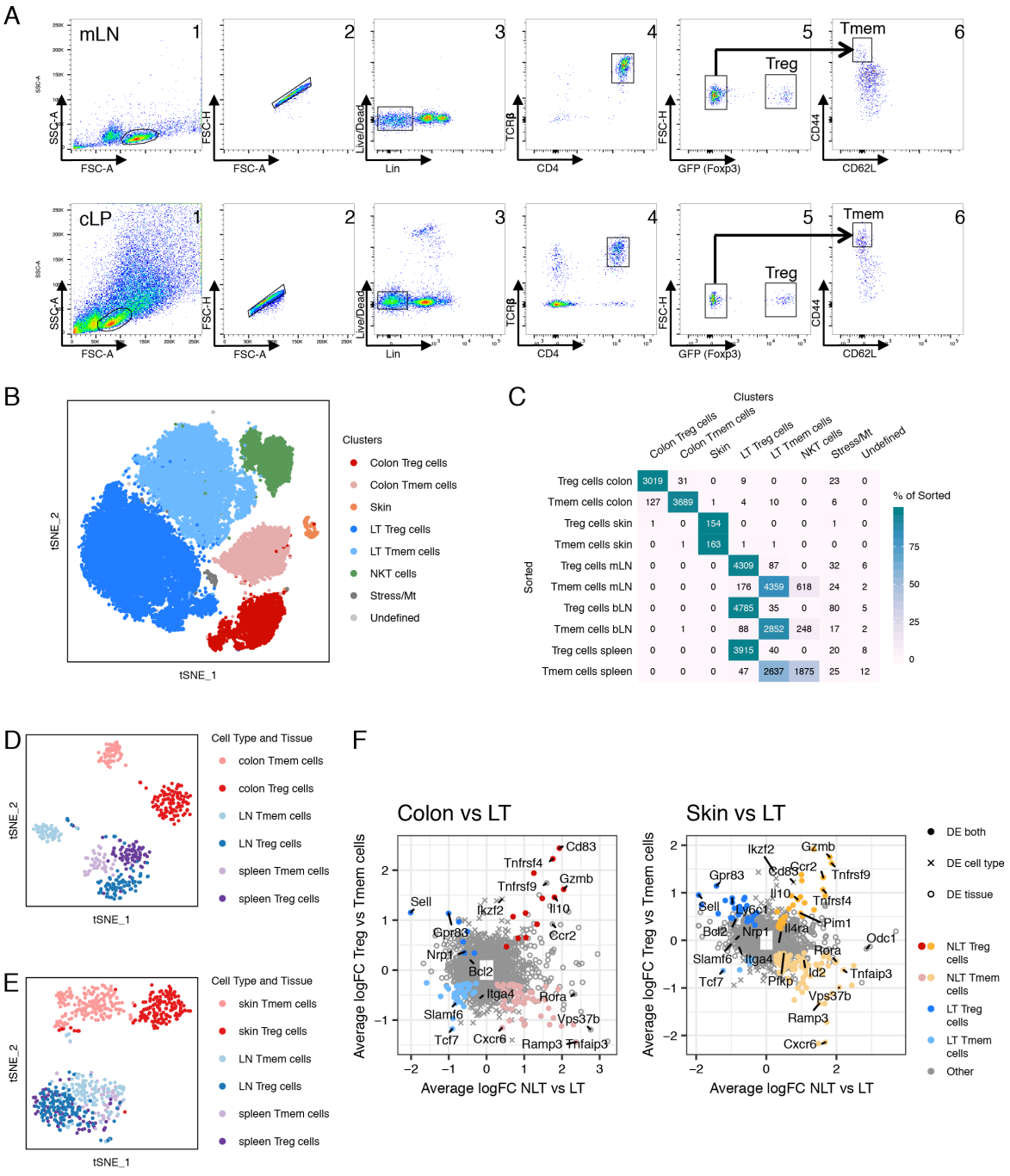


Fig. A.1: **Sorting and identification of Treg and Tmem cells (Related to Figure 2.1).**
(A) Flow cytometry-sorting strategy for sorting Treg and Tmem cells from (top) lymphoid (mLN) and (bottom) non-lymphoid (colonic lamina propria, cLP, as an example) organs. **(B)** tSNE projection of all 10x dataset cells passing QC, coloured by the resulting graph-based clustering. Cells from the NKT, Stress/Mt and Undefined clusters were removed from further analysis. (Continued on the following page.)

Fig. A.1: (continued) **(C)** Number of cells from each cluster in (D) originating from each sorted population. **(D and E)** Treg and Tmem cells were obtained with the same methodology as in Figure 2.1A, sequenced using Smart-seq2. t-SNE dimensionality reduction represents all sorted cells for each individual batch that passed quality control (see Methods). Colors match cell-type and tissue of origin. **(F)** Genes defining the identity of Treg and Tmem cells in lymphoid and non-lymphoid tissues, obtained from the Smart-seq2 datasets. Colon and skin were individually compared with their corresponding draining lymph node and spleen cells. Significantly expressed genes in each cell-type-tissue combination have an average log fold-change greater than 0.25 and and adjusted p-value lower than 0.05 (Wilcoxon test).

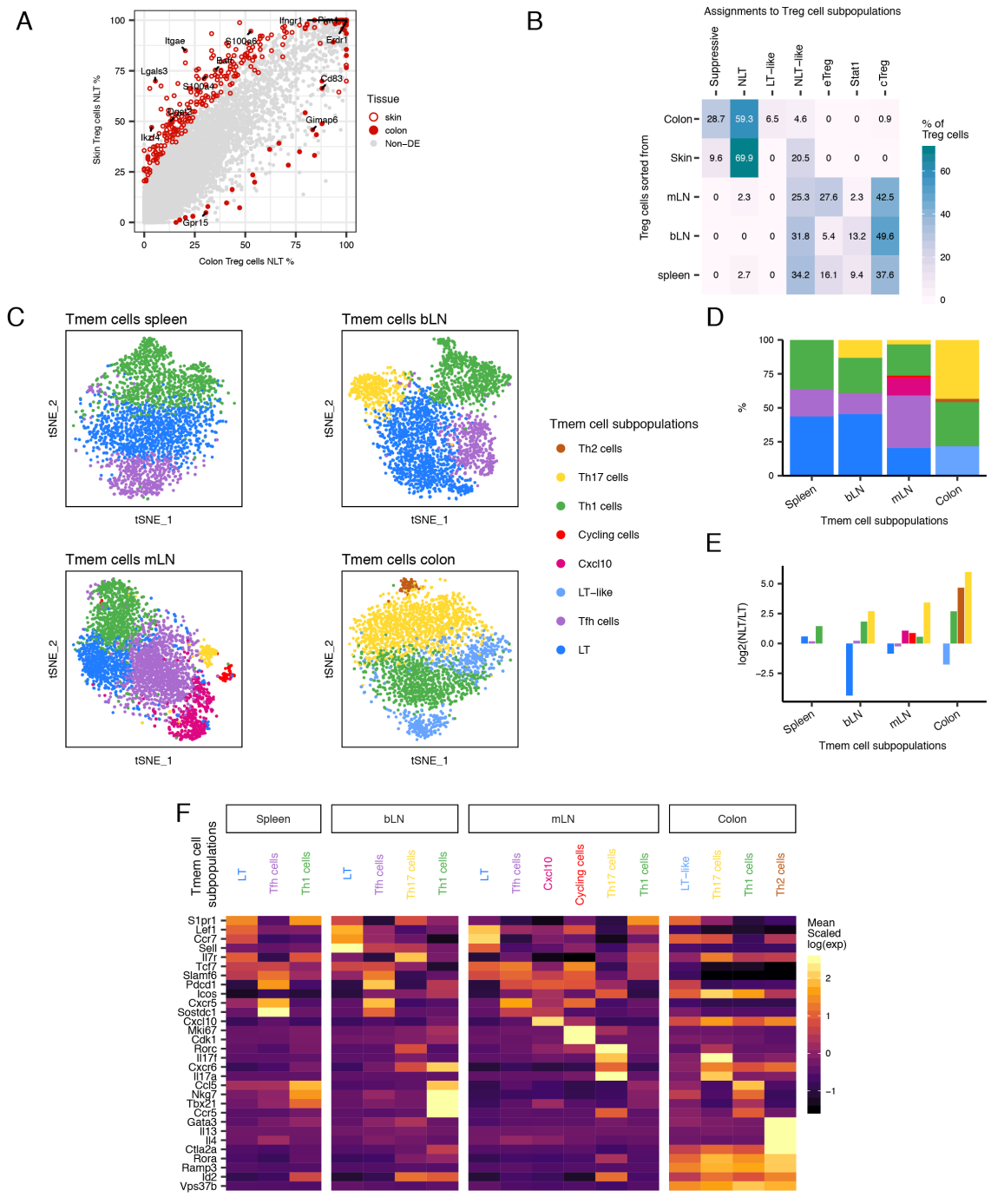


Fig. A.2: **Heterogeneity in SS2 and Tmem cell populations (Related to Figure 2.2).**
(A) Percentage of cells expressing each gene in skin Treg NLT and colon Treg NLT subpopulations in Smart-seq2 data. Genes that are upregulated in the skin Treg NLT subpopulation (log2(FC)>0.25 and adjusted p-value<0.05) are represented by an open circle, and genes upregulated in colon Treg NLT (log2(FC)<(-0.25) and adjusted p-value<0.05) are represented by a filled circle. (Continued on the following page.)

Fig. A.2: (continued) **(B)** Matching of Smart-seq2 Treg cells sorted populations to identified Treg subpopulations in the 10x dataset using a logistic regression model (85% accuracy, see Methods). Table shows the percentage of each sorted population (y-axis) that were labelled as each Treg cluster (x-axis). **(C)** t-SNE projection of Tmem cells per tissue coloured by subpopulations found using graph-based clustering. **(D)** Subpopulation marker gene mean expression levels (z-score) per subpopulation. Gene markers exhibit $|\log2(FC)| > 0.25$ and adjusted p-value<0.05 in the comparison of each subpopulation versus all the other cells within the same tissue. Values greater than 2.5 or lower than -1.5 are coloured equally. **(E)** Relative proportions of Tmem subpopulations within each tissue that revealed heterogeneity. **(F)** Measure of the NLT/LT signature score in each Tmem subpopulation, measured as the ratio between the number of NLT and LT genes that have been identified as significantly upregulated in each cluster.

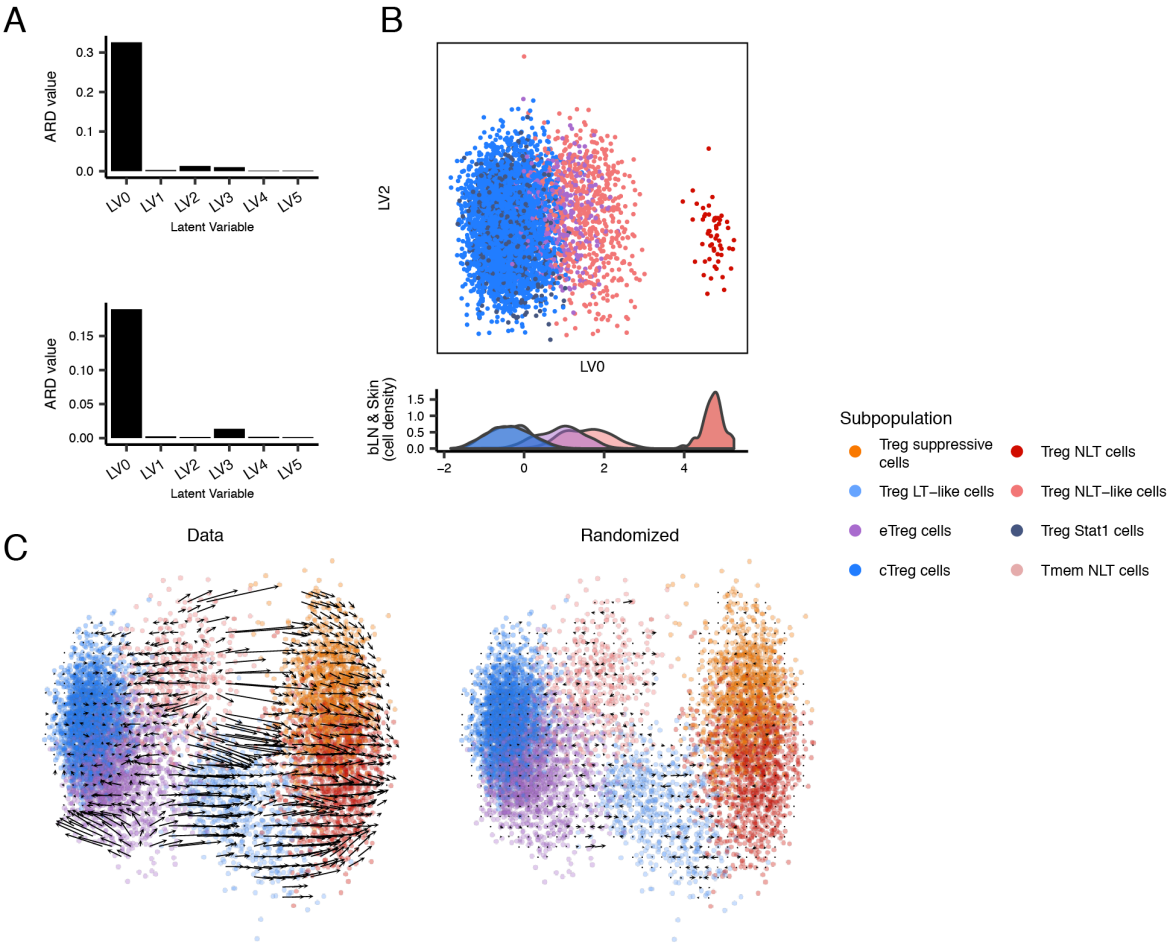


Fig. A.3: **Additional information on BGPLVM for the 10x dataset (Related to Figure 2.3)**
(A) Automatic Relevance Determination (ARD) plots for BGPLVM of Treg in mLN and colon (top, referring to Figure 2.3A), and bLN and skin (bottom, referring to Figure A.3B) datasets. These plots show the relevance of each latent variable extracted from the data. (B) BGPLVM dimensionality reduction of bLN and skin Treg cells from the 10X dataset (top), with a density plot showing the distribution along LV0 of each identified subpopulation (bottom). (C) Velocyto vectorfield overlaid on BGPLVM projection of mLN and colon Treg cells (related to Figure 2.3A).

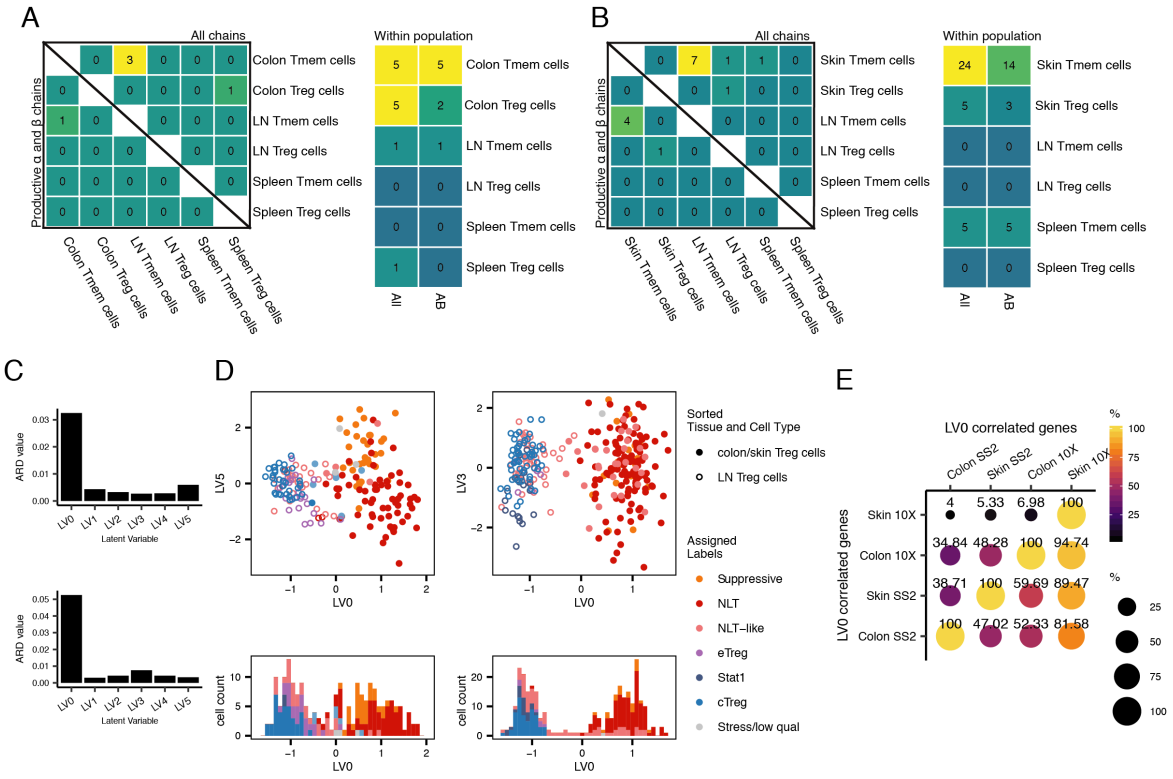


Fig. A.4: **Additional information on BGPLVM for the Smart-seq2 datasets (Related to Figure 2.3)**
(A and B) Clonotypes detected using TraCeR in the Smart-seq2 (A) Mouse Colon dataset, or (B) Mouse Skin dataset. In each panel, on the left, number of clonotypes detected spanning different tissues and cell type combinations. Top right half registers all events of TCR chain sharing, bottom left half only considers the sharing of productive α and β TCR chain, and on the right, number of clonotypes detected within each cell type and tissue, considering the sharing of any chain or productive α and β. **(C)** ARD plots for BGPLVM of Smart-seq2 Treg in mLN and colon (top, referring to panel D, left), and bLN and skin (bottom, referring to panel D, right) datasets. **(D)** BGPLVM dimensionality reduction of Smart-seq2 data of Treg from lymph nodes and non-lymphoid tissues (top), with a histogram plot showing the distribution along LV0 of each subpopulation identified (bottom). mLN and colon Treg are plotted on the left, while bLN and skin Treg are plotted on the right. Cells are coloured by the inferred subpopulation they belong to as per the predictions made in Figure A.2B. **(E)** Pairwise overlap between the sets of genes with absolute correlation with LV0 greater than 0.25 in each of the four steady-state datasets. The percentages refer to the proportion of the set on the x-axis that is overlapping the set on the y-axis.

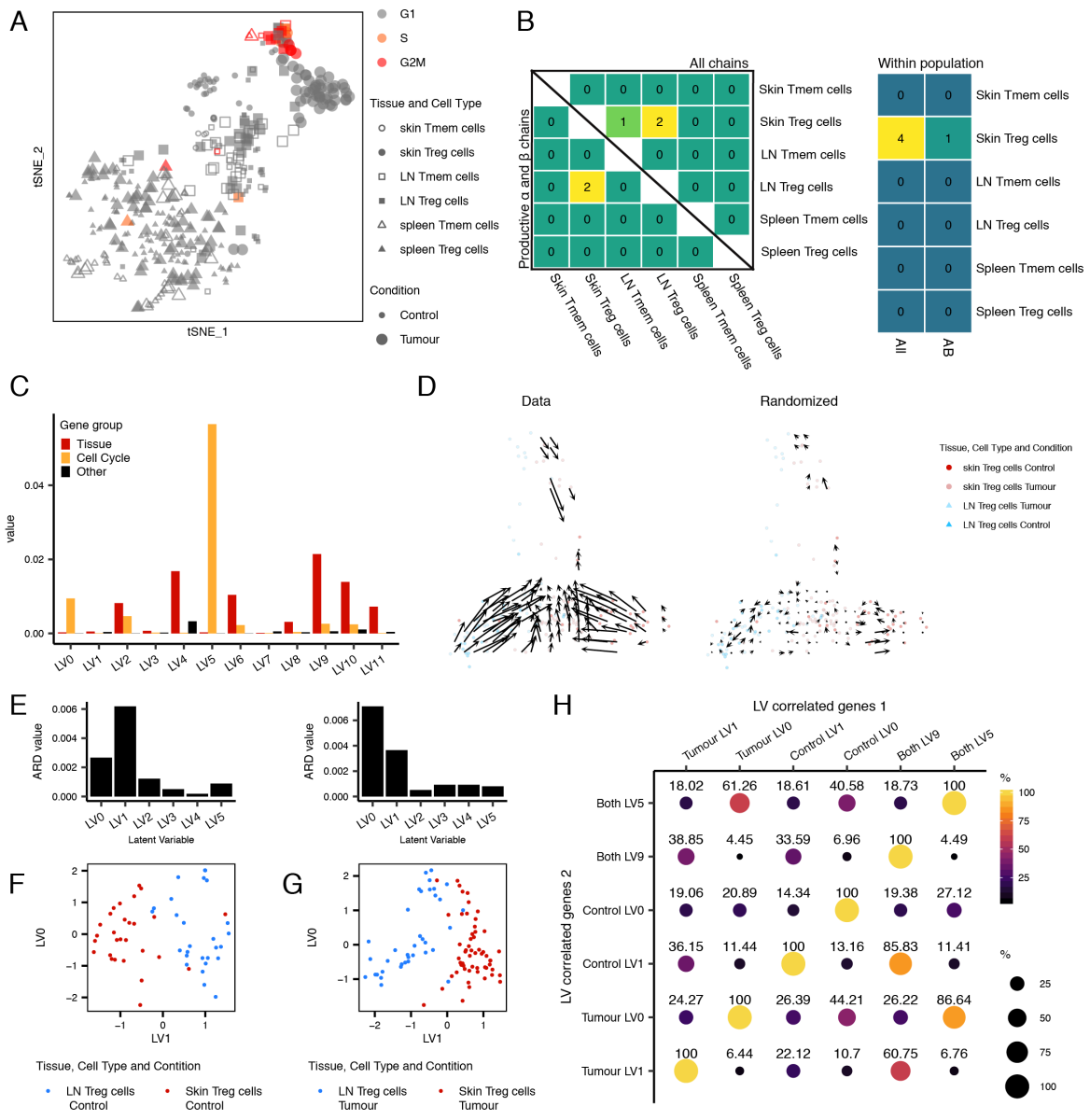


Fig. A.5: **Additional details on the MRD-BGPLVM projection (Related to Figure 2.4).**
(A) t-SNE dimensionality reduction coloured by cell cycle phase in the mouse melanoma dataset. **(B)** Clonotypes detected using TraCeR in the Mouse Melanoma dataset. On the left, number of clonotypes detected spanning different tissues and cell type combinations. Top right half registers all events of TCR chain sharing, bottom left half only considers the sharing of productive α and β TCR chain. On the right, number of clonotypes detected within each cell type and tissue, considering the sharing of any chain or productive α and β. **(C)** ARD plots for MRD-BGPLVM of Treg in control and melanoma conditions. Colours show effect of gene groups in each obtained latent variable. (Continued on the following page.)

Fig. A.5: (continued) **(D)** Velocyto vectorfield overlaid on MRD-BGPLVM projection of bLN and skin from both Control and Melanoma conditions (related to Figure 2.4D). **(E)** ARD plots for BGPLVM of Smart-seq2 Treg in bLN and skin in the Control condition (left, related to panel F), and Melanoma condition (bottom, related to panel G). **(F and G)** BGPLVM projection of bLN and skin in control (F) and melanoma (G) conditions, using the top two latent variables. **(H)** Pairwise overlap between the sets of genes with absolute correlation with LV0 greater than 0.25 in each subset of the melanoma dataset. The percentages refer to the proportion of the set on the x-axis that is overlapping the set on the y-axis.

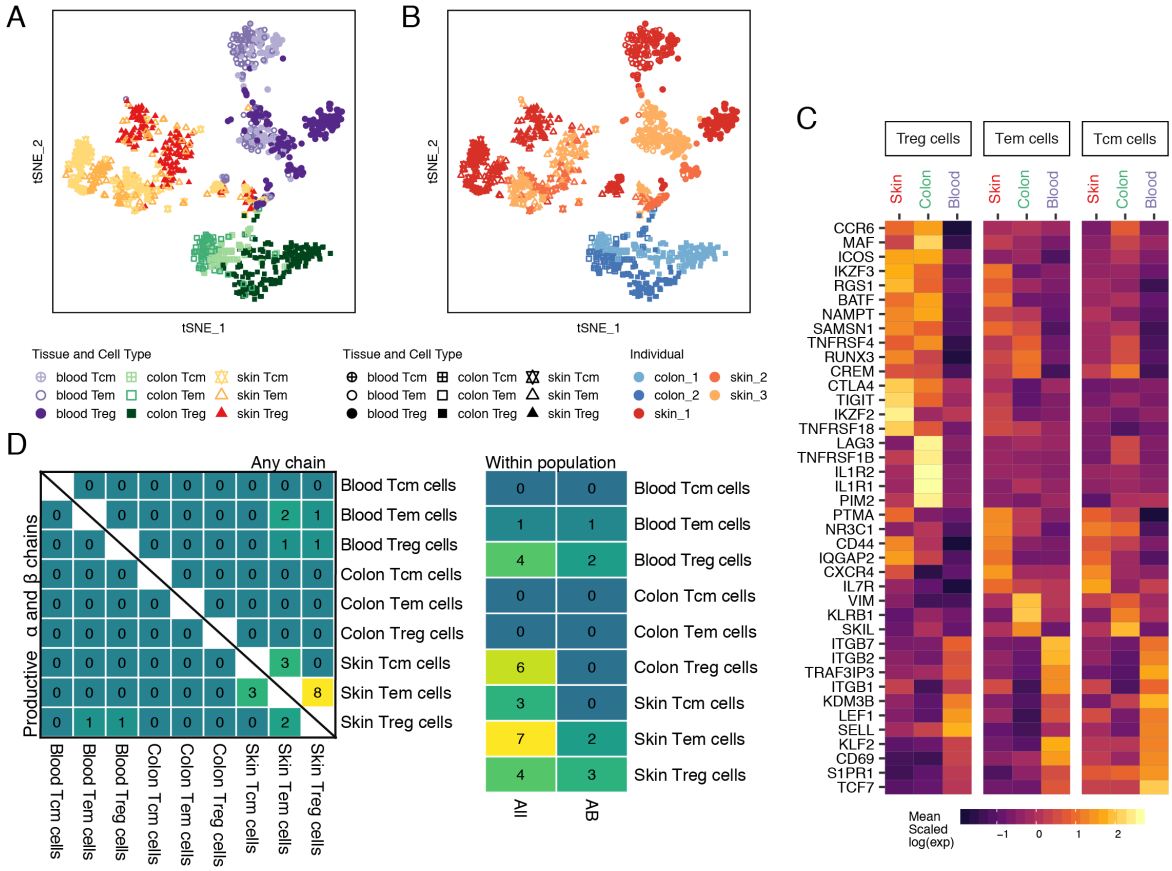


Fig. A.6: **Additional information on the Human dataset (Related to Figure 2.5). (A and B)** t-SNE dimensionality reduction. Shapes match cell type and tissue according to legend. Colours match either cell type and tissue (A) or sampled individual (B). **(C)** Z-score of mean expression levels of identified markers across all sampled cell types and tissues in human. **(D)** Clonotypes detected using TraCeR in the Human dataset. On the left, number of clonotypes detected spanning different tissues and cell type combinations. Top right half registers all events of TCR chain sharing, bottom left half only considers the sharing of productive α and β TCR chain. On the right, number of clonotypes detected within each cell type and tissue, considering the sharing of any chain or productive α and β.

A.3 Data and Code Accessibility

scRNA-seq data for this project has been deposited in ArrayExpress under the accession numbers E-MTAB-6072 and E-MTAB-7311. Processed data can be found in https://figshare.com/projects/Treg_scRNA-seq/38864, and analysis notebooks can be found in https://github.com/tomasgomes/Treg_analysis.

A.4 Full author list and contributions

Ricardo J. Miragaia*, Tomas Gomes*, Agnieszka Chomka, Laura Jardine, Angela Riedel, Ahmed N. Hegazy, Natasha Whibley, Andrea Tucci, Xi Chen, Ida Lindeman, Guy Emerton, Thomas Krausgruber, Jacqueline Shields, Muzlifah Haniffa, Fiona Powrie, Sarah A. Teichmann

*These authors contributed equally to this work

RJM, AC, AH, TK, FP and SAT conceived the project and designed steady-state experiments. RJM and AR designed the melanoma challenge experiments. AC and AH collected cells for steady-state mouse and human colonic datasets. LJ collected cells for human skin dataset. AR induced the melanoma challenge. RJM and AR collected cells for melanoma dataset. RJM performed scRNA-seq. TG, RJM and SAT planned the data analyses. TG and RJM analysed the data. IL performed the TraCeR analysis. RJM, TG, AC, SAT wrote the manuscript. SAT, FP, MH and JS supervised the work and edited manuscripts.

A.4.1 Acknowledgements

We thank V. Proserpio, M. Stubbington, T. Hagai, F.V. Braga, V. Svensson, J. Henriksson for helpful discussions and advice, T. Hagai, R. Vento-Tormo, K. Meyer, J. Park for critical reading of the manuscript, and B. Koppelman for editorial support. We thank WSI Single-cell Genomic Core Facility, WSI Sequencing Facility, WSI Flow Cytometry facilities, K. Polanski, as well as the Kennedy Institute Flow Cytometry Facility, for expert technical advice and assistance. RJM was supported by a PhD Fellowship from the Fundacao para a Ciencia e Tecnologia, Portugal (SFRH/BD/51950/2012), TG by the European Union's H2020 research and innovation programme "ENLIGHTTEN" under the Marie Sklodowska-Curie grant agreement 675395. This project was supported by ERC grants ThDEFINE and ThSWITCH. ANH was supported by

an EMBO long-term fellowship (ALTF 1161-2012) and a Marie Curie fellowship (PIEF-GA-2012-330621).

Appendix B

Additional information to Chapter 3

This Appendix contains supplementary figures for Chapter 3.

B.1 Supplementary Figures

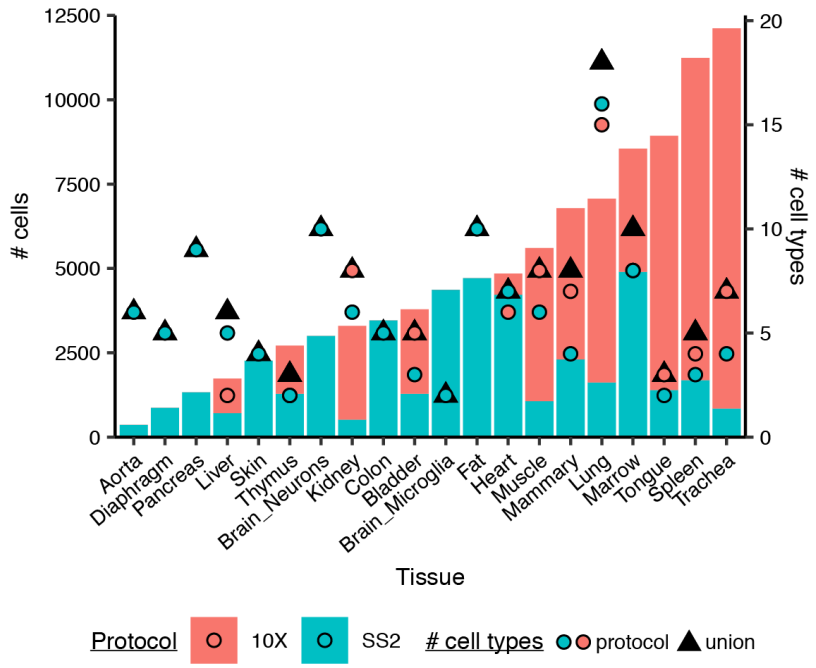


Fig. B.1: **Cell numbers in the *Tabula Muris* dataset**
Bars show number of cells (left y axis) collected from different tissues (x axis), split by scRNA-seq protocol (colour). Points show the number of cell types (right y axis) identified by protocol (coloured circles) or their union (triangle). 10X - Chromium (10X Genomics) protocol; SS2 - Smart-seq2 protocol.

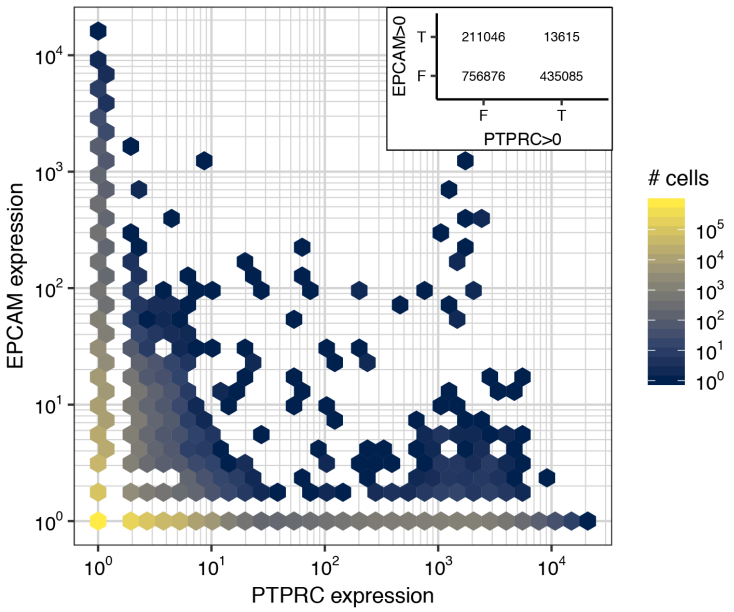


Fig. B.2: **Expression of *PTPRC* and *EPCAM* in human data collection (Related to Figure 3.6)**
2D-binned plot of single-cell expression of *PTPRC* (encoding for the CD45 receptor, an immune cell marker), and *EPCAM* (an epithelial cell marker). Inset table (top right) shows the number of cell expressing (T) or not (F) each of the genes. Cells expressing both genes are likely doublets or affected by ambient RNA in droplet-based experiments.

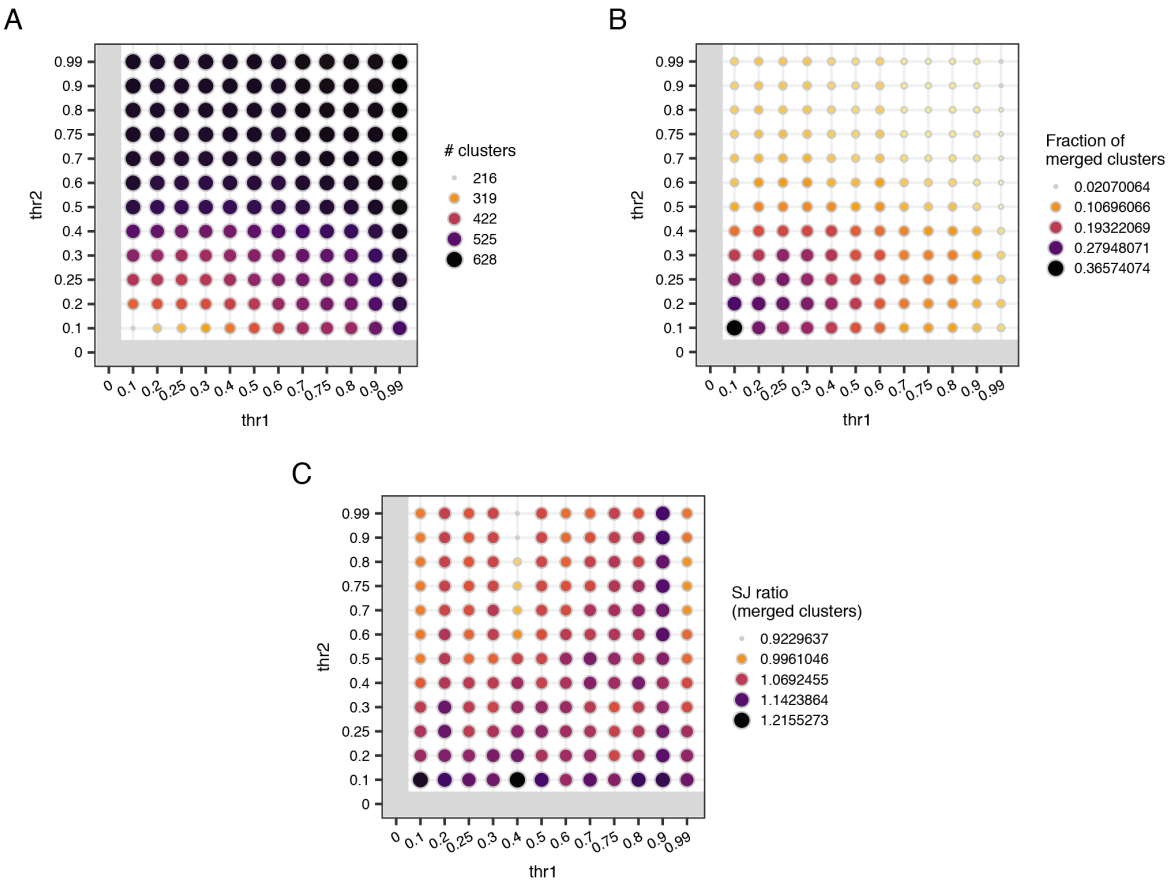


Fig. B.3: ***CellTypist* parameters grids with other statistics (Related to Figure 3.7)** Parameter grids for *CellTypist* showing variation in **(A)** total number of clusters; **(B)** fraction of merged clusters; **(C)** SJ ratio calculated only for merged clusters.

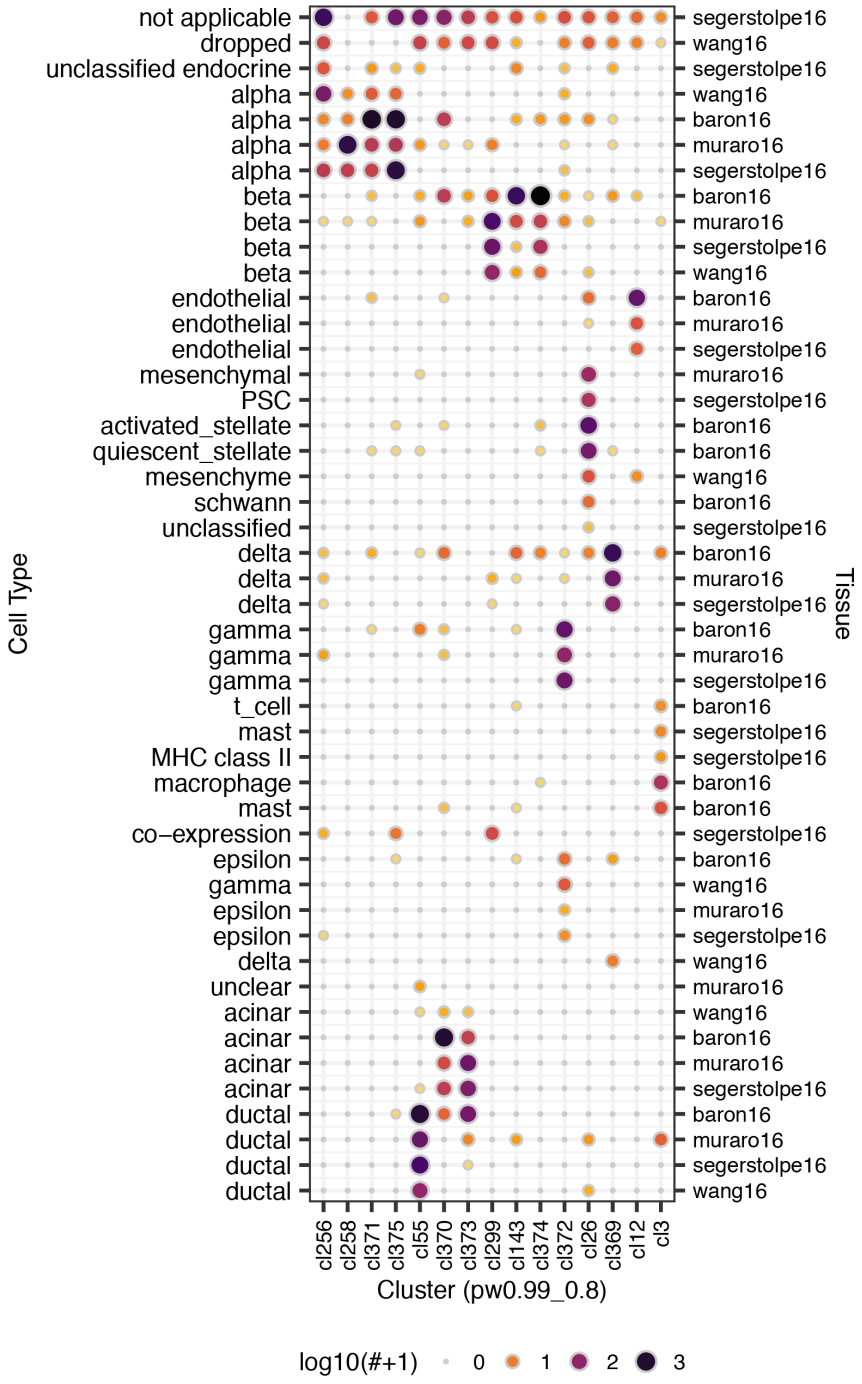


Fig. B.4: **Grouping of annotated cell types and datasets in human pancreas data (Related to Figure 3.7)**
Number of cells of each cluster coming from a specific dataset (right y-axis), with a particular cell type annotation (left y-axis). Pancreas was used for this example due to the consistent cell type annotations used across datasets. *CellTypist* parameters: thr1 = 0.99; thr2 = 0.8.

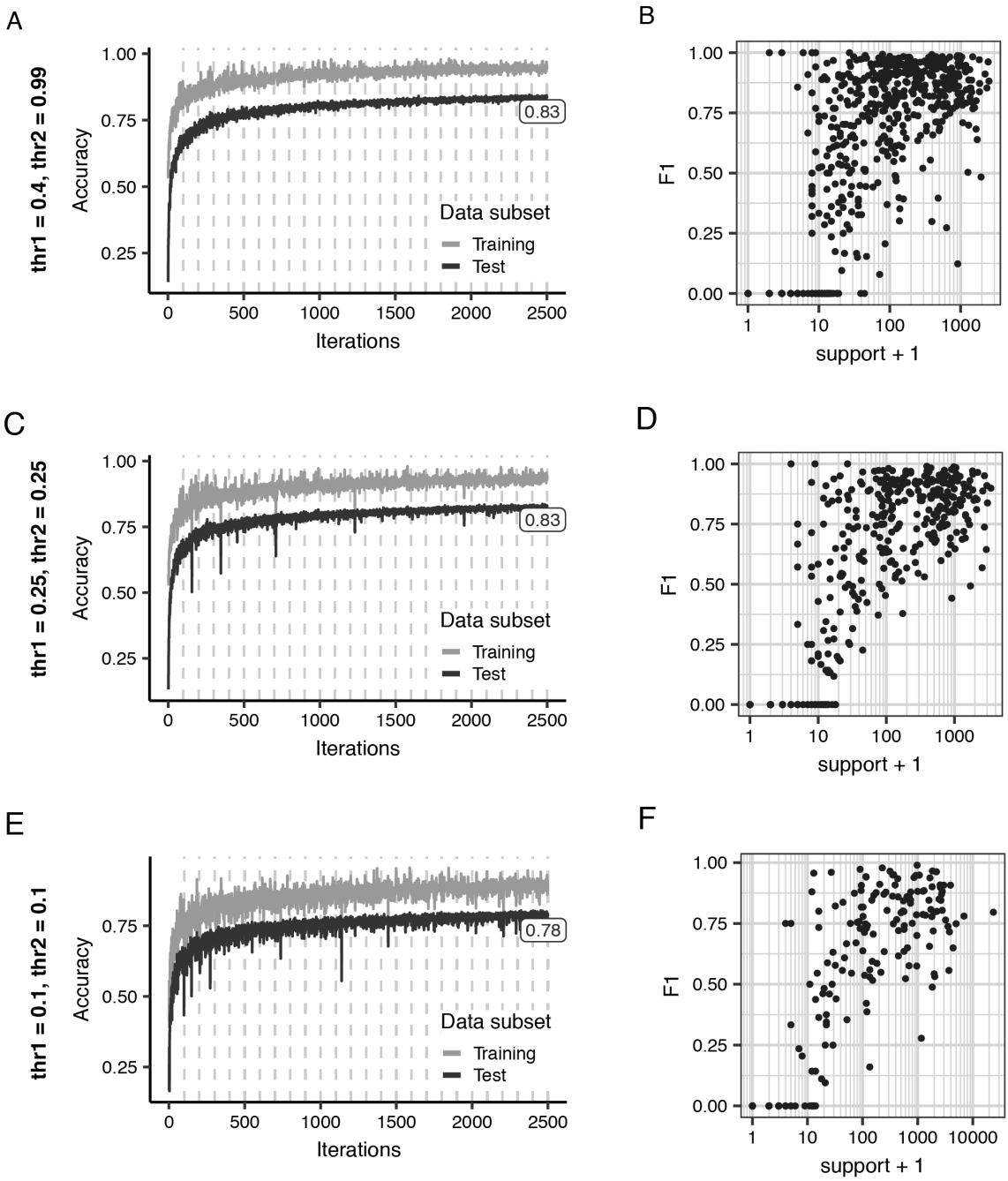


Fig. B.5: **Training statistics for other *CellTypist* models (Related to Figure 3.7)** For each model trained (thr1 = 0.4 and thr2 = 0.99 - top; thr1 = 0.25 and thr2 = 0.25 - middle; thr1 = 0.1 and thr2 = 0.1 - bottom): **(A, C, E)** accuracy during model fitting for training and held-out test data; **(B, D, F)** F1-score for each cluster label (black dots) as a function of class size (in log10 scale).

B.2 Supplementary Tables

Table B.1: F1 scores and class sizes for *CellTypist* trained on the *Tabula Muris* with cell type labels. Cell type labels were obtained from the annotation accompanying the *Tabula Muris* gene expression data, described in the original publication.

Cell Type	F1 Score	Support (Test set)	Total Cells
Bergmann glial cell	1.00	3	30
brain pericyte	1.00	13	132
Brush cell of epithelium proper of large intestine	1.00	4	45
enteroendocrine cell	1.00	2	25
mesothelial cell	1.00	3	26
neuronal stem cell	1.00	4	36
pancreatic ductal cell	1.00	13	131
type II pneumocyte	1.00	18	183
microglial cell	1.00	433	4329
keratinocyte stem cell	1.00	137	1371
oligodendrocyte	1.00	119	1186
basal cell	0.99	167	1668
luminal epithelial cell of mammary gland	0.99	55	552
type B pancreatic cell	0.99	41	411
chondroblast	0.99	38	380
B cell	0.98	1237	12382
kidney tubule cell	0.98	218	2182
mesenchymal cell	0.98	184	1842
stromal cell	0.98	1261	12610
skeletal muscle satellite stem cell	0.98	44	442
neuron	0.98	20	196
oligodendrocyte precursor cell	0.97	20	202
mesenchymal stem cell of adipose	0.97	192	1924
basal cell of epidermis	0.97	652	6520
hematopoietic stem cell	0.97	267	2672
hepatocyte	0.97	141	1405
skeletal muscle satellite cell	0.97	90	895
pancreatic A cell	0.97	29	287
epithelial cell	0.97	102	1017
astrocyte of the cerebral cortex	0.96	40	403
Fraction A pre-pro B cell	0.96	24	240
endocardial cell	0.96	24	240
T cell	0.96	835	8346
keratinocyte	0.95	278	2777
epithelial cell of large intestine	0.95	179	1793
endothelial cell	0.95	692	6914
large intestine goblet cell	0.95	81	814

Table B.2: F1 scores and class sizes for *CellTypist* trained on the *Tabula Muris* with cell type labels. Cell type labels were obtained from the annotation accompanying the *Tabula Muris* gene expression data, described in the original publication. (continued)

Cell Type	F1 Score	Support (Test set)	Total Cells
fibroblast	0.95	248	2487
pancreatic D cell	0.95	9	91
pancreatic acinar cell	0.94	18	177
enterocyte of epithelium of large intestine	0.94	78	782
luminal cell of lactiferous duct	0.93	43	430
neutrophil	0.93	82	820
mesenchymal stem cell	0.93	163	1630
endothelial cell of hepatic sinusoid	0.92	20	196
epidermal cell	0.92	45	445
granulocyte	0.91	156	1559
monocyte	0.91	106	1056
neuroendocrine cell	0.90	54	543
Kupffer cell	0.89	5	51
fenestrated cell	0.88	41	414
cardiac muscle cell	0.87	22	223
smooth muscle cell	0.87	37	367
natural killer cell	0.85	117	1168
macrophage	0.85	194	1924
leukocyte	0.84	187	1878
bladder cell	0.84	146	1455
erythrocyte	0.81	21	208
ciliated cell	0.80	5	55
ciliated epithelial cell	0.80	2	20
pancreatic stellate cell	0.80	3	29
myeloid cell	0.79	53	527
pancreatic PP cell	0.78	11	107
kidney collecting duct cell	0.76	12	116
stem cell of epidermis	0.75	4	45
dendritic cell	0.71	43	438
unknown	0.67	63	625
Clara cell	0.67	2	18
hematopoietic cell	0.67	2	17
mast cell	0.44	2	22
basal cell of urothelium	0.40	36	365
basal cell of epithelium of trachea	0.12	4	36
epicardial adipocyte	0.00	9	93
lung neuroendocrine cell	0.00	0	2
type I pneumocyte	0.00	0	2

Table B.3: F1 scores and class sizes for *CellTypist* trained on the *Tabula Muris* with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3.

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl103	1.00	21	207	endothelial cell (93.7%)	Lung (100%)
cl124	1.00	8	82	neuron (100%)	Brain_Neurons (100%)
cl135	1.00	3	26	neuron (100%)	Brain_Neurons (100%)
cl138	1.00	6	57	neuron (98.2%)	Brain_Neurons (100%)
cl147	1.00	15	147	kidney collecting duct cell (57.1%)	Kidney (100%)
cl154	1.00	2	17	neuron (100%)	Brain_Neurons (100%)
cl166	1.00	5	50	type B pancreatic cell (100%)	Pancreas (100%)
cl175	1.00	10	96	Fraction A pre-pro B cell (65.6%)	Marrow (100%)
cl185	1.00	5	47	Brush cell of epithelium proper of large intestine (89.4%)	Colon (100%)
cl186	1.00	1	8	neuron (100%)	Brain_Neurons (100%)
cl189	1.00	46	462	hepatocyte (97.4%)	Liver (100%)
cl193	1.00	4	39	unknown (61.5%)	Aorta (100%)
cl194	1.00	1	14	enteroendocrine cell (100%)	Colon (100%)
cl20	1.00	12	123	pancreatic ductal cell (99.2%)	Pancreas (100%)
cl27	1.00	1	9	enteroendocrine cell (100%)	Colon (100%)
cl65	1.00	2	21	pancreatic stellate cell (90.5%)	Pancreas (100%)
cl73	1.00	2	21	leukocyte (100%)	Pancreas (100%)
cl89	1.00	1	13	astrocyte of the cerebral cortex (69.2%)	Brain_Neurons (69.2%)
cl161	0.99	96	959	keratinocyte (93.8%)	Tongue (100%)
cl150	0.99	83	833	keratinocyte stem cell (97.7%)	Skin (100%)
cl70	0.99	163	1631	basal cell (99.8%)	Mammary (100%)
cl157	0.99	68	682	luminal epithelial cell of mammary gland (57.8%)	Mammary (100%)
cl190	0.99	58	583	hepatocyte (100%)	Liver (100%)
cl143	0.99	54	544	kidney tubule cell (98%)	Kidney (100%)
cl171	0.99	54	539	keratinocyte stem cell (99.6%)	Skin (100%)
cl26	0.99	36	357	luminal cell of lactiferous duct (51%)	Mammary (100%)
cl81	0.99	72	715	endothelial cell (93.3%)	Lung (100%)
cl38	0.98	223	2227	fibroblast (99.3%)	Heart (100%)
cl100	0.98	105	1052	mesenchymal cell (97.1%)	Bladder (100%)
cl108	0.98	49	494	epithelial cell (54.5%)	Trachea (55.9%)
cl31	0.98	496	4961	stromal cell (97.4%)	Trachea (100%)
cl49	0.98	82	823	mesenchymal cell (98.4%)	Bladder (100%)
cl122	0.97	39	386	stromal cell (99.7%)	Lung (100%)
cl37	0.97	367	3665	stromal cell (98.4%)	Trachea (100%)
cl17	0.97	74	742	epithelial cell of large intestine (99.9%)	Colon (100%)
cl95	0.97	90	899	T cell (99.8%)	Thymus (100%)
cl6	0.97	227	2267	hematopoietic stem cell (73.7%)	Marrow (100%)
cl165	0.97	101	1008	kidney tubule cell (98.5%)	Kidney (100%)
cl47	0.97	116	1157	bladder cell (78%)	Bladder (100%)
cl1	0.97	454	4543	basal cell of epidermis (95%)	Tongue (100%)
cl54	0.97	180	1798	granulocyte (61.5%)	Marrow (100%)
cl136	0.97	63	631	T cell (100%)	Thymus (100%)
cl24	0.97	63	631	epithelial cell (92.4%)	Trachea (100%)
cl106	0.96	41	407	chondroblast (93.1%)	Muscle (99.3%)
cl180	0.96	39	392	kidney tubule cell (98.7%)	Kidney (100%)
cl141	0.96	128	1279	skeletal muscle satellite cell (67.9%)	Muscle (70%)
cl71	0.96	24	239	smooth muscle cell (98.7%)	Heart (100%)
cl32	0.96	24	238	mesenchymal stem cell (97.5%)	Diaphragm (100%)
cl85	0.96	82	822	natural killer cell (99.8%)	Lung (100%)
cl44	0.96	143	1434	mesenchymal stem cell (93.4%)	Muscle (100%)
cl176	0.96	24	242	enterocyte of epithelium of large intestine (97.9%)	Colon (100%)
cl181	0.96	23	229	astrocyte ofthe cerebral cortex (92.6%)	Brain_Neurons (100%)

Table B.4: F1 scores and class sizes for *CellTypist* trained on the *Tabula Muris* with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. (continued 1)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl76	0.95	23	232	endocardial cell (98.7%)	Heart (100%)
cl151	0.95	32	324	hepatocyte (99.7%)	Liver (100%)
cl126	0.95	22	223	keratinocyte (95.5%)	Tongue (100%)
cl39	0.95	159	1593	endothelial cell (72.8%)	Trachea (65.3%)
cl99	0.95	87	868	stromal cell (99.1%)	Lung (100%)
cl140	0.95	68	684	hematopoietic stem cell (64%)	Marrow (100%)
cl83	0.95	38	379	macrophage (53.6%)	Marrow (98.2%)
cl187	0.95	19	188	type II pneumocyte (97.3%)	Lung (100%)
cl159	0.94	17	174	oligodendrocyte precursor cell (99.4%)	Brain_Neurons (100%)
cl62	0.94	196	1955	endothelial cell (98.7%)	Heart (68%)
cl86	0.94	101	1012	basal cell of epidermis (95.6%)	Tongue (99.4%)
cl125	0.94	17	168	type B pancreatic cell (98.2%)	Pancreas (100%)
cl144	0.94	77	771	basal cell of epidermis (95.2%)	Tongue (100%)
cl64	0.94	150	1504	T cell (99.1%)	Mammary (100%)
cl48	0.94	68	678	stromal cell (99%)	Mammary (100%)
cl131	0.94	86	862	oligodendrocyte (96.6%)	Brain_Neurons (100%)
cl60	0.94	239	2391	T cell (99.3%)	Spleen (100%)
cl74	0.94	753	7534	B cell (98.6%)	Spleen (70.5%)
cl14	0.94	123	1232	B cell (60%)	Mammary (100%)
cl110	0.94	63	630	bladder cell (81.4%)	Bladder (99.4%)
cl18	0.94	97	965	leukocyte (97.3%)	Trachea (100%)
cl96	0.94	73	729	mesenchymal stem cell of adipose (55.4%)	Fat (55.7%)
cl55	0.94	132	1318	endothelial cell (98%)	Muscle (100%)
cl113	0.94	143	1434	keratinocyte (97.6%)	Tongue (99.7%)
cl132	0.94	23	226	epithelial cell of large intestine (83.2%)	Colon (100%)
cl43	0.93	120	1197	monocyte (69.2%)	Marrow (100%)
cl112	0.93	8	84	large intestine goblet cell (100%)	Colon (100%)
cl97	0.93	8	80	mesenchymal stem cell of adipose (100%)	Fat (100%)
cl173	0.93	22	220	epidermal cell (93.2%)	Skin (100%)
cl67	0.93	29	294	granulocyte (94.6%)	Fat (100%)
cl129	0.93	36	360	stromal cell (99.2%)	Lung (100%)
cl0	0.93	118	1182	T cell (87.9%)	Thymus (100%)
cl127	0.93	35	348	large intestine goblet cell (87.9%)	Colon (100%)
cl35	0.92	48	481	leukocyte (88.4%)	Heart (100%)
cl82	0.92	18	179	brain pericyte (58.1%)	Brain_Neurons (58.1%)
cl21	0.92	31	307	endothelial cell (94.8%)	Mammary (100%)
cl57	0.92	54	537	B cell (98.7%)	Muscle (100%)
cl11	0.91	11	115	ciliated cell (47%)	Lung (100%)
cl58	0.91	21	211	endothelial cell of hepatic sinusoid (85.8%)	Liver (100%)
cl121	0.90	41	413	epithelial cell of large intestine (99.8%)	Colon (100%)
cl42	0.90	20	204	neuroendocrine cell (95.1%)	Trachea (100%)
cl88	0.90	31	308	myeloid cell (99%)	Fat (100%)
cl52	0.90	54	538	endothelial cell (99.3%)	Fat (100%)
cl184	0.89	12	117	pancreatic acinar cell (97.4%)	Pancreas (100%)
cl34	0.89	33	327	macrophage (96.9%)	Muscle (100%)
cl130	0.89	18	176	large intestine goblet cell (96%)	Colon (100%)
cl72	0.89	94	938	mesenchymal stem cell of adipose (99.9%)	Fat (100%)
cl94	0.89	62	617	stromal cell (97.4%)	Lung (100%)
cl142	0.88	16	156	type B pancreatic cell (100%)	Pancreas (100%)
cl77	0.88	31	309	leukocyte (48.9%)	Lung (58.3%)
cl13	0.88	33	326	unknown (61%)	Muscle (100%)
cl145	0.88	38	384	basal cell of epidermis (80.7%)	Skin (100%)

Table B.5: F1 scores and class sizes for *CellTypist* trained on the *Tabula Muris* with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. (continued 2)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl45	0.88	55	545	stromal cell (94.3%)	Lung (100%)
cl102	0.87	17	170	stromal cell (87.6%)	Lung (97.1%)
cl153	0.87	14	137	pancreatic A cell (71.5%)	Pancreas (100%)
cl84	0.86	22	216	monocyte (81.9%)	Lung (100%)
cl15	0.86	24	239	macrophage (74.9%)	Kidney (100%)
cl51	0.86	123	1229	microglial cell (99.6%)	Brain_Microglia (100%)
cl172	0.86	18	177	large intestine goblet cell (52%)	Colon (100%)
cl87	0.86	29	294	dendritic cell (88.4%)	Lung (100%)
cl90	0.86	8	76	endothelial cell (100%)	Aorta (100%)
cl10	0.85	271	2706	B cell (98.1%)	Spleen (100%)
cl7	0.85	53	535	macrophage (63.6%)	Spleen (100%)
cl177	0.85	21	208	enterocyte of epithelium of large intestine (90.4%)	Colon (100%)
cl53	0.83	27	273	endothelial cell (96.3%)	Lung (100%)
cl183	0.83	13	128	large intestine goblet cell (98.4%)	Colon (100%)
cl29	0.83	170	1700	microglial cell (100%)	Brain_Microglia (100%)
cl162	0.83	24	237	enterocyte of epithelium of large intestine (98.7%)	Colon (100%)
cl115	0.83	22	219	mesenchymal stem cell of adipose (99.5%)	Fat (100%)
cl188	0.83	20	197	astrocyte of the cerebral cortex (87.3%)	Brain_Neurons (100%)
cl119	0.83	32	317	oligodendrocyte (99.7%)	Brain_Neurons (100%)
cl40	0.83	26	260	neuroendocrine cell (74.2%)	Trachea (100%)
cl107	0.82	9	94	epidermal cell (43.6%)	Skin (96.8%)
cl109	0.82	26	265	mesenchymal stem cell of adipose (100%)	Fat (100%)
cl12	0.81	143	1427	microglial cell (98.5%)	Brain_Microglia (100%)
cl152	0.81	15	155	pancreatic A cell (98.1%)	Pancreas (100%)
cl56	0.80	47	467	T cell (72.8%)	Fat (100%)
cl9	0.80	5	49	epithelial cell (95.9%)	Fat (100%)
cl69	0.80	11	113	neuroendocrine cell (90.3%)	Trachea (100%)
cl80	0.79	22	222	neutrophil (98.2%)	Fat (100%)
cl36	0.79	16	162	myeloid cell (90.1%)	Fat (100%)
cl149	0.79	19	186	epidermal cell (57.5%)	Skin (100%)
cl68	0.79	77	766	T cell (74.7%)	Marrow (60.8%)
cl191	0.78	13	131	epithelial cell of large intestine (59.5%)	Colon (100%)
cl63	0.77	31	315	T cell (91.1%)	Lung (100%)
cl2	0.77	16	156	Kupffer cell (32.7%)	Liver (100%)
cl158	0.77	25	255	epithelial cell of large intestine (76.1%)	Colon (100%)
cl197	0.76	10	97	unknown (58.8%)	Brain_Neurons (100%)
cl114	0.75	35	354	macrophage (99.4%)	Lung (100%)
cl28	0.75	9	94	B cell (69.1%)	Diaphragm (100%)
cl101	0.75	5	48	endothelial cell (93.8%)	Fat (91.7%)
cl111	0.75	5	53	oligodendrocyte (64.2%)	Brain_Neurons (100%)
cl4	0.71	31	313	cardiac muscle cell (62.9%)	Heart (100%)
cl5	0.70	21	210	fibroblast (66.2%)	Kidney (100%)
cl148	0.70	11	107	pancreatic D cell (57.9%)	Pancreas (100%)
cl30	0.69	36	362	T cell (96.4%)	Muscle (100%)
cl93	0.67	1	9	unknown (44.4%)	Aorta (100%)
cl41	0.62	19	187	macrophage (73.8%)	Lung (100%)
cl25	0.62	6	60	leukocyte (90%)	Bladder (100%)
cl8	0.59	11	106	unknown (67.9%)	Brain_Neurons (100%)
cl50	0.57	5	46	fibroblast (58.7%)	Aorta (100%)
cl33	0.55	8	78	endothelial cell (92.3%)	Diaphragm (100%)
cl19	0.50	1	14	leukocyte (78.6%)	Pancreas (100%)
cl79	0.50	3	26	smooth muscle cell (96.2%)	Fat (100%)
cl16	0.48	7	69	endothelial cell (85.5%)	Bladder (100%)

Table B.6: F1 scores and class sizes for *CellTypist* trained on the *Tabula Muris* with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. (continued 3)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl66	0.45	26	262	B cell (98.1%)	Lung (100%)
cl59	0.44	4	36	leukocyte (77.8%)	Kidney (100%)
cl23	0.22	7	67	epicardial adipocyte (47.8%)	Aorta (100%)
cl155	0.18	2	18	pancreatic acinar cell (100%)	Pancreas (100%)
cl104	0.00	0	5	mesenchymal stem cell of adipose (100%)	Fat (100%)
cl105	0.00	0	4	mesenchymal stem cell of adipose (100%)	Fat (100%)
cl116	0.00	0	5	endothelial cell (100%)	Fat (100%)
cl117	0.00	1	12	smooth muscle cell (100%)	Brain_Neurons (100%)
cl118	0.00	0	3	epithelial cell of large intestine (66.7%)	Colon (100%)
cl120	0.00	0	5	endothelial cell (100%)	Brain_Neurons (100%)
cl123	0.00	2	16	pancreatic PP cell (100%)	Pancreas (100%)
cl128	0.00	1	6	pancreatic PP cell (100%)	Pancreas (100%)
cl133	0.00	1	9	oligodendrocyte precursor cell (100%)	Brain_Neurons (100%)
cl134	0.00	0	4	epithelial cell of large intestine (100%)	Colon (100%)
cl137	0.00	1	14	pancreatic A cell (64.3%)	Pancreas (100%)
cl139	0.00	3	34	pancreatic A cell (64.7%)	Pancreas (100%)
cl146	0.00	1	9	T cell (88.9%)	Fat (100%)
cl156	0.00	3	26	pancreatic D cell (100%)	Pancreas (100%)
cl160	0.00	1	8	type B pancreatic cell (100%)	Pancreas (100%)
cl163	0.00	0	3	type B pancreatic cell (100%)	Pancreas (100%)
cl164	0.00	0	4	neuron (100%)	Brain_Neurons (100%)
cl167	0.00	0	3	pancreatic ductal cell (100%)	Pancreas (100%)
cl168	0.00	1	8	smooth muscle cell (37.5%)	Aorta (100%)
cl169	0.00	1	6	unknown (100%)	Brain_Neurons (100%)
cl170	0.00	0	4	pancreatic A cell (100%)	Pancreas (100%)
cl174	0.00	1	6	fibroblast (66.7%)	Aorta (100%)
cl178	0.00	0	3	epicardial adipocyte (100%)	Aorta (100%)
cl179	0.00	1	6	type B pancreatic cell (100%)	Pancreas (100%)
cl182	0.00	0	3	Brush cell of epithelium proper of large intestine (100%)	Colon (100%)
cl192	0.00	0	5	epithelial cell of large intestine (60%)	Colon (100%)
cl195	0.00	0	4	pancreatic acinar cell (100%)	Pancreas (100%)
cl196	0.00	5	50	pancreatic acinar cell (78%)	Pancreas (100%)
cl22	0.00	5	53	skeletal muscle satellite stem cell (90.6%)	Diaphragm (100%)
cl3	0.00	1	6	keratinocyte stem cell (100%)	Skin (100%)
cl46	0.00	1	11	epicardial adipocyte (54.5%)	Aorta (100%)
cl61	0.00	2	16	B cell (93.8%)	Diaphragm (100%)
cl75	0.00	1	10	hematopoietic cell (60%)	Aorta (60%)
cl78	0.00	2	20	hematopoietic cell (55%)	Aorta (55%)
cl91	0.00	0	3	endothelial cell (100%)	Aorta (100%)
cl92	0.00	7	70	endothelial cell (58.6%)	Aorta (100%)
cl98	0.00	0	3	macrophage (100%)	Diaphragm (100%)

Table B.7: Human scRNA-seq datasets collected and corresponding cell numbers

Dataset	**Reference**	**# cells**
baron16	(Baron et al., 2016)	8.569
bjorklund16	(Bjorklund et al., 2016)	648
gierahn17	(Gierahn et al., 2017)	3.694
guo18	(Guo et al., 2018)	12.053
habib17	(Habib et al., 2017)	14.963
hcaImmune18	HCA Data Portal	593.844
henry18	(Henry et al., 2018)	109.061
jaitin19	(Jaitin et al., 2019)	13.199
james20	*Unpublished*	32.228
lamanno16	(La Manno et al., 2016)	1.977
li19	(Li et al., 2019b)	1.886
masuda19	(Masuda et al., 2019)	6.144
menon18	(Menon et al., 2018)	9.846
miragaia18	(Miragaia et al., 2019)	1.168
muraro16	(Muraro et al., 2016)	2.126
nowakowski17	(Nowakowski et al., 2017)	4.261
popescu19	(Popescu et al., 2019)	113.063
segal19	(Segal et al., 2019)	1.475
segerstolpe16	(Segerstolpe et al., 2016)	3.363
smillie19	(Smillie et al., 2019)	110.110
sohni19	(Sohni et al., 2019)	34.729
takeda19	(Takeda et al., 2019)	33.257
vento18	(Vento-Tormo et al., 2018)	69.883
vieira19	(Braga et al., 2019)	26.013
wang16	(Wang et al., 2016)	635
young18	(Young et al., 2018)	44.526
zhang18	(Zhang et al., 2018)	5.989
zheng17	(Zheng et al., 2017)	163.234
Total		1.421.944

Table B.8: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it.

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl19	1.00	6	56	No annotation (100%)	Intestine (100%)
cl198	1.00	4	41	No annotation (100%)	Brain_Microglia (100%)
cl264	1.00	8	76	Endo (m) (96.1%)	Decidua (100%)
cl311	1.00	12	123	Smooth muscle (56.9%)	Lung Parenchyma (100%)
cl319	1.00	1	7	SCT (100%)	Placenta (100%)
cl362	1.00	2	21	No annotation (100%)	Brain_Microglia (100%)
cl67	1.00	33	326	Endo L (97.5%)	Decidua (100%)
cl307	0.99	392	3916	Type 2 (97.8%)	Lung Parenchyma (100%)
cl282	0.99	450	4496	Myoid cells (6.6%)	Testis (100%)
cl295	0.99	46	464	dS1 (35.3%)	Decidua (100%)
cl376	0.99	1039	10393	No annotation (100%)	Prostate (100%)
cl242	0.99	905	9047	dS1 (50.8%)	Decidua (100%)
cl34	0.99	71	710	Fibroblasts (99.7%)	Lung Parenchyma (100%)
cl179	0.99	1013	10132	dNK2 (49%)	Decidua (100%)
cl131	0.98	193	1932	Macrophages (97%)	Lung Parenchyma (100%)
cl451	0.98	976	9763	CD19+ B (96.4%)	Blood (100%)
cl83	0.98	310	3104	Leydig cells (18.8%)	Testis (100%)
cl12	0.98	30	302	endothelial (90.4%)	Pancreas (100%)
cl27	0.98	369	3685	Endothelial cells (8.8%)	Testis (100%)
cl266	0.98	55	547	Neutrophils (74.2%)	Lung Parenchyma (100%)
cl127	0.98	26	258	Macrophages (98.1%)	Lung Parenchyma (100%)
cl269	0.98	26	264	NK (87.1%)	Lung Parenchyma (100%)
cl55	0.98	180	1797	ductal (88.1%)	Pancreas (100%)
cl44	0.98	551	5513	dM1 (50.2%)	Decidua (100%)
cl286	0.98	173	1731	fFB1 (99.2%)	Placenta (100%)
cl97	0.98	188	1875	dP1 (54.7%)	Decidua (100%)
cl24	0.98	91	910	Macrophages (37.1%)	Testis (100%)
cl109	0.98	424	4240	No annotation (100%)	BoneMarrow (100%)
cl102	0.98	353	3529	fFB1 (0.1%)	Testis (99.8%)
cl122	0.98	128	1276	HB (98.2%)	Placenta (100%)
cl345	0.98	101	1012	MGE newborn neurons (29.5%)	Brain (100%)
cl64	0.97	78	783	Endo (m) (97.3%)	Decidua (100%)
cl494	0.97	314	3137	No annotation (100%)	BoneMarrow (100%)
cl133	0.97	79	791	Macrophages (97.1%)	Lung Parenchyma (100%)
cl35	0.97	89	885	dM3 (86.4%)	Placenta (100%)
cl284	0.97	53	533	Ciliated (99.6%)	Upper airway (100%)
cl369	0.97	89	889	delta (95.6%)	Pancreas (100%)
cl79	0.97	702	7018	NK (52.3%)	Liver (100%)
cl23	0.97	18	178	Neutrophils (52.8%)	Upper airway (100%)
cl442	0.97	894	8944	CD56+ NK (91.2%)	Blood (100%)
cl341	0.97	227	2269	Sperm (84.5%)	Testis (100%)
cl113	0.97	2447	24471	CD19+ B (0.3%)	Blood (100%)
cl252	0.97	228	2276	Macrophages (90%)	Lung Parenchyma (100%)
cl447	0.97	919	9192	Treg (0%)	Blood (100%)
cl192	0.97	155	1551	Neutrophils (93.5%)	Lung Parenchyma (100%)
cl523	0.97	985	9851	No annotation (100%)	BoneMarrow (100%)
cl338	0.97	15	145	EVT (82.8%)	Decidua (100%)
cl379	0.97	116	1159	Type 2 (97.7%)	Lung Parenchyma (100%)
cl281	0.97	99	994	dS3 (85.5%)	Decidua (100%)
cl517	0.96	595	5952	No annotation (100%)	BoneMarrow (100%)
cl75	0.96	162	1619	No annotation (100%)	BoneMarrow (100%)
cl118	0.96	89	887	Macrophages (98.9%)	Lung Parenchyma (100%)
cl327	0.96	266	2658	Differentiating Spermatogonia (11.6%)	Testis (100%)
cl378	0.96	419	4194	No annotation (100%)	Prostate (100%)
cl312	0.96	141	1413	Early Primary Spermatocytes (38.7%)	Testis (100%)
cl314	0.96	396	3959	No annotation (100%)	Prostate (100%)
cl321	0.96	76	759	dNK1 (31.4%)	Decidua (100%)

Table B.9: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 1)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl208	0.96	134	1335	dM1 (51%)	Decidua (99.7%)
cl69	0.96	1184	11840	CD19+ B (30.9%)	Blood (100%)
cl322	0.96	107	1070	Elongated Spermatids (66%)	Testis (100%)
cl570	0.96	70	699	OPC (86.6%)	Brain (100%)
cl326	0.96	365	3653	No annotation (100%)	Kidney (100%)
cl22	0.96	12	121	No annotation (100%)	Brain_Microglia (100%)
cl455	0.95	1072	10721	CD8+/CD45RA+ Naive Cytotoxic (0.9%)	Blood (100%)
cl49	0.95	244	2441	Fibroblast (29.9%)	Liver (100%)
cl543	0.95	172	1722	No annotation (100%)	BoneMarrow (100%)
cl380	0.95	543	5433	No annotation (100%)	Prostate (100%)
cl7	0.95	102	1015	No annotation (100%)	Omentum Adipose Tissue (100%)
cl15	0.95	49	486	MG (70.2%)	Brain (100%)
cl11	0.95	57	572	Endothelium (65%)	Lung Parenchyma (56.3%)
cl606	0.95	623	6230	WNT2B+ Fos-lo 1 (27.4%)	Colon (100%)
cl413	0.95	1834	18344	CD8+/CD45RA+ Naive Cytotoxic (0.1%)	Blood (100%)
cl45	0.95	486	4862	Macrophages (70.2%)	Colon (100%)
cl308	0.95	66	656	Type 2 (98.6%)	Lung Parenchyma (100%)
cl40	0.95	267	2669	Kupffer Cell (19.8%)	Liver (100%)
cl440	0.95	931	9305	CD34+ (87.6%)	Blood (100%)
cl316	0.94	122	1216	Secretory (91.2%)	Upper airway (100%)
cl77	0.94	269	2694	No annotation (100%)	BoneMarrow (100%)
cl155	0.94	237	2373	pro-B cell (25%)	Liver (99.8%)
cl240	0.94	74	744	dNK2 (47.6%)	Decidua (100%)
cl2	0.94	461	4613	No annotation (100%)	Prostate (100%)
cl503	0.94	692	6923	No annotation (100%)	BoneMarrow (100%)
cl268	0.94	176	1760	dS1 (94%)	Decidua (100%)
cl243	0.94	81	808	dS1 (87.5%)	Decidua (100%)
cl63	0.94	41	406	Secretory (67.5%)	Lung Parenchyma (100%)
cl582	0.94	193	1933	Th cell (0.1%)	Kidney (100%)
cl261	0.94	108	1078	EVT (97%)	Placenta (100%)
cl206	0.94	44	439	fFB1 (98.9%)	Placenta (100%)
cl283	0.94	165	1650	dT CD8 (26.2%)	Decidua (88.3%)
cl403	0.94	357	3568	Megakaryocyte (45.3%)	Liver (100%)
cl68	0.94	207	2071	ILC precursor (39.3%)	Liver (100%)
cl492	0.93	343	3425	No annotation (100%)	BoneMarrow (100%)
cl370	0.93	124	1239	acinar (80%)	Pancreas (100%)
cl372	0.93	63	630	gamma (87.6%)	Pancreas (100%)
cl456	0.93	1151	11506	CD34+ (1.4%)	Blood (100%)
cl47	0.93	493	4926	Normal_cell (4%)	Kidney (100%)
cl377	0.93	432	4324	No annotation (100%)	Prostate (100%)
cl536	0.93	30	296	No annotation (100%)	BoneMarrow (100%)
cl547	0.93	60	595	No annotation (100%)	BoneMarrow (100%)
cl219	0.93	57	574	Endothelial cells (3.8%)	Testis (100%)
cl540	0.93	179	1785	No annotation (100%)	BoneMarrow (100%)
cl432	0.93	743	7430	No annotation (100%)	Prostate (100%)
cl375	0.93	207	2065	alpha (90.4%)	Pancreas (100%)
cl515	0.93	638	6377	No annotation (100%)	BoneMarrow (100%)
cl374	0.93	204	2040	beta (98.9%)	Pancreas (100%)
cl271	0.93	105	1047	No annotation (100%)	Omentum Adipose Tissue (100%)
cl371	0.93	134	1343	alpha (97.5%)	Pancreas (100%)
cl401	0.92	455	4553	No annotation (100%)	Prostate (100%)
cl506	0.92	914	9140	No annotation (100%)	BoneMarrow (100%)
cl504	0.92	1595	15946	No annotation (100%)	BoneMarrow (100%)
cl260	0.92	81	811	Basal (98.3%)	Upper airway (100%)
cl250	0.92	135	1353	No annotation (100%)	axLN (100%)

Table B.10: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 2)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl301	0.92	14	138	Secretory (87%)	Upper airway (100%)
cl542	0.92	175	1748	No annotation (100%)	BoneMarrow (100%)
cl490	0.92	1166	11658	No annotation (100%)	BoneMarrow (100%)
cl568	0.92	761	7609	TA 1 (37.4%)	Colon (100%)
cl458	0.92	1177	11771	CD56+ NK (47.2%)	Blood (100%)
cl474	0.92	92	920	Mast cell (49.1%)	Liver (100%)
cl48	0.92	18	177	Tcm (57.1%)	Skin (100%)
cl495	0.92	308	3084	No annotation (100%)	BoneMarrow (100%)
cl509	0.92	760	7595	No annotation (100%)	BoneMarrow (100%)
cl277	0.92	12	115	No annotation (100%)	Brain_Microglia (100%)
cl470	0.92	1345	13446	NK CD16+ (9.1%)	Blood (100%)
cl592	0.92	925	9250	No annotation (100%)	BoneMarrow (100%)
cl340	0.92	259	2593	Hepatocyte (40.9%)	Liver (100%)
cl335	0.92	57	566	Late primary Spermatocytes (41.3%)	Testis (100%)
cl255	0.92	114	1144	EVT (98.8%)	Placenta (100%)
cl70	0.91	381	3813	No annotation (100%)	BoneMarrow (100%)
cl59	0.91	36	356	END (66%)	Brain (100%)
cl233	0.91	407	4066	VCT (99.9%)	Placenta (100%)
cl511	0.91	573	5731	No annotation (100%)	BoneMarrow (100%)
cl464	0.91	610	6102	Mid Erythroid (25.5%)	Liver (100%)
cl323	0.91	107	1068	Spermatogonial Stem cell (0.7%)	Testis (100%)
cl554	0.91	498	4982	B cell IgA plasma (45.9%)	Colon (100%)
cl552	0.91	93	926	No annotation (100%)	BoneMarrow (100%)
cl258	0.91	75	747	alpha (99.9%)	Pancreas (100%)
cl457	0.91	2216	22163	PB Naive CD4 (0.1%)	Blood (100%)
cl25	0.91	95	951	No annotation (100%)	Omentum Adipose Tissue (100%)
cl38	0.91	32	319	Unknown1 (24.8%)	Brain (100%)
cl232	0.91	70	702	CD4 Tfh (71.2%)	mLN (100%)
cl486	0.90	441	4414	No annotation (100%)	BoneMarrow (100%)
cl404	0.90	1304	13039	No annotation (100%)	Prostate (100%)
cl612	0.90	698	6977	Immature Enterocytes 1 (35.6%)	Colon (100%)
cl214	0.90	16	157	ILC2 (83.4%)	Tonsil (100%)
cl581	0.90	279	2792	Normal_cell (6.5%)	Kidney (100%)
cl488	0.90	389	3888	No annotation (100%)	BoneMarrow (100%)
cl293	0.90	42	424	Ciliated (99.8%)	Upper airway (100%)
cl115	0.90	277	2767	Kupffer Cell (26.7%)	Liver (100%)
cl373	0.90	63	632	acinar (59.3%)	Pancreas (100%)
cl429	0.90	567	5673	No annotation (100%)	Prostate (100%)
cl518	0.90	234	2344	No annotation (100%)	BoneMarrow (100%)
cl56	0.90	575	5748	Mid Erythroid (29.5%)	Liver (100%)
cl417	0.90	257	2570	No annotation (100%)	Prostate (100%)
cl491	0.90	366	3663	No annotation (100%)	BoneMarrow (100%)
cl505	0.90	1430	14301	No annotation (100%)	BoneMarrow (100%)
cl278	0.90	100	1000	VCT (97.7%)	Placenta (100%)
cl390	0.90	55	547	earlyRG (2.2%)	Brain (100%)
cl575	0.90	278	2780	Endothelium; Mixed_phenotype (<0.1%)	Kidney (100%)
cl279	0.90	85	849	VCT (93.1%)	Placenta (100%)
cl36	0.89	172	1717	Endothelial (25.6%)	Colon (100%)
cl210	0.89	104	1038	dNK3 (87.5%)	Decidua (100%)
cl444	0.89	2513	25130	CD8+/CD45RA+ Naive Cytotoxic (59%)	Blood (100%)
cl299	0.89	77	774	beta (89%)	Pancreas (100%)

Table B.11: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 3)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl538	0.89	204	2038	No annotation (100%)	BoneMarrow (100%)
cl430	0.89	680	6797	No annotation (100%)	Prostate (100%)
cl567	0.89	789	7894	TA 1 (75.4%)	Colon (100%)
cl498	0.89	276	2763	No annotation (100%)	BoneMarrow (100%)
cl569	0.89	490	4899	Plasma (94.8%)	Colon (100%)
cl247	0.89	4	37	No annotation (100%)	Omentum Adipose Tissue (100%)
cl259	0.89	15	154	EVT (100%)	Placenta (100%)
cl336	0.89	10	101	No annotation (100%)	Brain_Microglia (100%)
cl88	0.89	47	466	Endothelium (1.3%)	axLN (98.7%)
cl478	0.89	1857	18568	CD4+/CD45RA+/CD25- Naive T (48.1%)	Blood (100%)
cl57	0.89	157	1571	DCs (60.2%)	Lung Parenchyma (100%)
cl8	0.89	225	2246	dT CD8 (31.5%)	Decidua (100%)
cl512	0.88	679	6788	No annotation (100%)	BoneMarrow (100%)
cl100	0.88	17	169	Plasma (1.8%)	Omentum Adipose Tissue (98.2%)
cl267	0.88	13	126	No annotation (100%)	Brain_Microglia (100%)
cl433	0.88	821	8207	No annotation (100%)	Prostate (100%)
cl408	0.88	73	727	EVT (97.2%)	Placenta (100%)
cl1	0.88	724	7237	Renal_cell_carcinoma (3.4%)	Kidney (100%)
cl339	0.88	192	1917	Spermatogonial Stem cell (4.5%)	Testis (100%)
cl392	0.88	56	563	Newborn Excitatory Neuron - late born (1.2%)	Brain (100%)
cl317	0.88	123	1226	VCT (97.5%)	Placenta (100%)
cl280	0.88	9	92	No annotation (100%)	Brain_Microglia (100%)
cl382	0.88	8	83	Fibroblasts (77.1%)	Lung Parenchyma (100%)
cl391	0.87	115	1147	Newborn Excitatory Neuron - early born (42%)	Brain (100%)
cl134	0.87	946	9463	Early Erythroid (44.5%)	Liver (100%)
cl617	0.87	176	1764	CD69+ Mast (50.6%)	Colon (100%)
cl350	0.87	19	190	Unclassified (75.3%)	Brain (100%)
cl355	0.87	43	432	Ciliated (97.5%)	Upper airway (100%)
cl18	0.87	21	206	ILC3 (96.6%)	Tonsil (100%)
cl387	0.87	268	2680	Spermatogonial Stem cell (6.7%)	Testis (100%)
cl256	0.87	80	801	not applicable (65.9%)	Pancreas (99.6%)
cl74	0.87	605	6048	No annotation (100%)	Prostate (100%)
cl508	0.87	864	8640	No annotation (100%)	BoneMarrow (100%)
cl610	0.87	310	3098	CD4+ Memory (81.7%)	Colon (100%)
cl46	0.87	601	6013	CD14+ Monocyte (30%)	Blood (100%)
cl431	0.87	742	7415	No annotation (100%)	Prostate (100%)
cl613	0.87	273	2734	Follicular (74.8%)	Colon (100%)
cl4	0.86	27	272	Sertoli cells (5.9%)	Testis (100%)
cl309	0.86	33	325	No annotation (100%)	axLN (100%)
cl500	0.86	453	4529	No annotation (100%)	BoneMarrow (100%)
cl14	0.86	87	865	T cell (76.2%)	Lung Parenchyma (100%)
cl510	0.86	689	6892	No annotation (100%)	BoneMarrow (100%)
cl13	0.86	24	240	No annotation (100%)	hnLN (100%)
cl203	0.86	11	108	fFB2 (97.2%)	Placenta (100%)
cl37	0.86	8	79	NK (88.6%)	Tonsil (100%)
cl62	0.86	7	72	No annotation (100%)	Brain_Microglia (100%)
cl501	0.86	464	4637	No annotation (100%)	BoneMarrow (100%)
cl601	0.86	84	835	B cell IgA plasma (71.9%)	Colon (100%)
cl590	0.85	386	3857	B cell IgA plasma (70.7%)	Colon (100%)
cl343	0.85	97	968	NSC (13.4%)	Brain (100%)

Table B.12: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 4)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl516	0.85	1394	13939	No annotation (100%)	BoneMarrow (100%)
cl238	0.85	141	1411	B cell memory (40%)	mLN (100%)
cl595	0.85	637	6372	Plasma (97.3%)	Colon (100%)
cl611	0.85	168	1677	B cell IgA plasma (64.9%)	Colon (100%)
cl306	0.85	38	384	Myoid cells (5.5%)	Testis (100%)
cl475	0.84	146	1457	Mid Erythroid (24.6%)	Liver (100%)
cl101	0.84	221	2213	Plasma (85.6%)	Colon (100%)
cl497	0.84	278	2784	No annotation (100%)	BoneMarrow (100%)
cl51	0.84	834	8335	Kupffer Cell (27.8%)	Liver (100%)
cl438	0.84	314	3141	HSC_MPP (36%)	Liver (100%)
cl452	0.84	1198	11980	MO (0.6%)	Blood (100%)
cl183	0.84	49	494	Mast cell (86.2%)	Lung Parenchyma (86.6%)
cl534	0.84	951	9505	No annotation (100%)	BoneMarrow (100%)
cl514	0.84	938	9378	No annotation (100%)	BoneMarrow (100%)
cl93	0.84	528	5279	Myeloid (31.8%)	Blood (100%)
cl110	0.84	132	1318	No annotation (100%)	BoneMarrow (100%)
cl576	0.83	1353	13527	Th cell (3.7%)	Kidney (100%)
cl386	0.83	62	619	MGE Progenitors (22%)	Brain (100%)
cl43	0.83	7	66	No annotation (100%)	Intestine (100%)
cl439	0.83	1849	18491	Bcell (0%)	Blood (100%)
cl220	0.83	234	2342	Sperm (0.6%)	Testis (100%)
cl26	0.83	69	692	activated_stellate (40.5%)	Pancreas (100%)
cl507	0.83	892	8923	No annotation (100%)	BoneMarrow (100%)
cl275	0.83	16	161	No annotation (100%)	Prostate (100%)
cl50	0.83	32	316	No annotation (100%)	Omentum Adipose Tissue (100%)
cl332	0.83	94	937	GABA1 (50.6%)	Brain (100%)
cl428	0.83	42	421	No annotation (100%)	Prostate (100%)
cl615	0.83	252	2520	CD4+ CD25- T cells (31.3%)	Colon (100%)
cl520	0.83	214	2137	No annotation (100%)	BoneMarrow (100%)
cl270	0.83	84	842	No annotation (100%)	axLN (100%)
cl329	0.83	253	2534	Private (4.9%)	Kidney (100%)
cl493	0.82	329	3289	No annotation (100%)	BoneMarrow (100%)
cl30	0.82	9	93	ILC1 (100%)	Tonsil (100%)
cl276	0.82	35	349	No annotation (100%)	Brain_Microglia (100%)
cl304	0.82	361	3610	Sertoli cells (0.1%)	Testis (100%)
cl551	0.82	107	1070	No annotation (100%)	BoneMarrow (100%)
cl549	0.82	116	1156	No annotation (100%)	BoneMarrow (100%)
cl418	0.82	979	9790	Mid Erythroid (67.6%)	Liver (100%)
cl389	0.82	17	171	No annotation (100%)	Brain_Microglia (100%)
cl96	0.82	21	205	EVT (94.6%)	Placenta (98.5%)
cl246	0.82	71	714	Basal (99.4%)	Upper airway (100%)
cl3	0.82	13	125	macrophage (43.2%)	Pancreas (100%)
cl448	0.82	921	9207	CD14+ Monocyte (1.8%)	Blood (100%)
cl87	0.82	642	6419	Treg NL-like (0%)	BoneMarrow (100%)
cl465	0.82	292	2919	Mid Erythroid (71.9%)	Liver (100%)
cl229	0.82	35	347	No annotation (100%)	hnLN (100%)
cl71	0.82	1018	10179	CD8+ Cytotoxic T (49.7%)	Blood (100%)
cl524	0.81	15	148	No annotation (100%)	BoneMarrow (100%)
cl480	0.81	291	2914	CD4+ CD25high T cells (28.1%)	Blood (100%)
cl263	0.81	62	617	Basal (99.2%)	Upper airway (100%)
cl586	0.81	88	881	exPFC1 (55.2%)	Brain (100%)
cl477	0.81	155	1554	Early Erythroid (55.1%)	Liver (100%)
cl537	0.81	53	525	No annotation (100%)	BoneMarrow (100%)
cl143	0.81	62	620	beta (90.6%)	Pancreas (100%)
cl577	0.81	83	834	NK cell 1 (4.2%)	Kidney (100%)
cl178	0.81	40	401	No annotation (100%)	hnLN (100%)
cl502	0.80	571	5707	No annotation (100%)	BoneMarrow (100%)

Table B.13: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 5)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl571	0.80	459	4593	TA 1 (51.9%)	Colon (100%)
cl298	0.80	15	146	No annotation (100%)	Brain_Microglia (100%)
cl553	0.80	86	861	No annotation (100%)	BoneMarrow (100%)
cl359	0.80	31	309	Unclassified (66%)	Brain (100%)
cl360	0.80	458	4577	Sertoli cells (0%)	Testis (100%)
cl76	0.80	1487	14872	MO (3.5%)	Blood (100%)
cl80	0.80	61	614	No annotation (100%)	Omentum Adipose Tissue (100%)
cl296	0.80	57	565	Early Born Deep Layer/ subplate Excitatory Neuron V1 (25%)	Brain (100%)
cl460	0.80	487	4866	PB Naive CD4 (0.1%)	Blood (100%)
cl459	0.79	1360	13598	CD8+/CD45RA+ Naive Cytotoxic (58.9%)	Blood (100%)
cl272	0.79	58	579	No annotation (100%)	axLN (100%)
cl227	0.79	121	1205	No annotation (100%)	axLN (100%)
cl41	0.79	25	249	EVT (87.1%)	Decidua (100%)
cl358	0.79	494	4938	Sertoli cells (0.1%)	Testis (100%)
cl262	0.79	86	862	Basal (98.5%)	Upper airway (100%)
cl89	0.79	471	4714	CD8+ LP (58%)	Colon (100%)
cl160	0.79	166	1656	No annotation (100%)	Omentum Adipose Tissue (99.8%)
cl313	0.79	53	528	SCT (78.6%)	Placenta (100%)
cl449	0.78	1453	14529	CD4+/CD25 T Reg (49.5%)	Blood (100%)
cl626	0.78	117	1172	B cell IgA plasma (53.2%)	Colon (100%)
cl616	0.78	249	2494	CD8+ IELs (52.4%)	Colon (100%)
cl361	0.78	112	1124	exPFC1 (88.9%)	Brain (100%)
cl399	0.78	50	499	GABA2 (55.5%)	Brain (100%)
cl435	0.78	1197	11966	Kupffer Cell (67%)	Liver (100%)
cl254	0.77	19	190	No annotation (100%)	hnLN (100%)
cl422	0.77	412	4117	Late Erythroid (23.9%)	Liver (100%)
cl618	0.77	158	1583	Plasma (84.2%)	Colon (100%)
cl556	0.77	548	5477	Cycling TA (47.1%)	Colon (100%)
cl60	0.77	114	1137	No annotation (100%)	Omentum Adipose Tissue (100%)
cl402	0.77	497	4972	No annotation (100%)	Prostate (100%)
cl318	0.77	159	1591	VCT (97.4%)	Placenta (100%)
cl453	0.76	983	9833	CD4+/CD25 T Reg (35.3%)	Blood (100%)
cl548	0.76	1017	10174	No annotation (100%)	BoneMarrow (100%)
cl483	0.76	73	726	Sertoli cells (0.3%)	Testis (100%)
cl400	0.76	774	7742	Mid Erythroid (63.6%)	Liver (100%)
cl412	0.76	58	581	Kupffer Cell (32.7%)	Liver (100%)
cl388	0.76	59	590	exCA3 (62.5%)	Brain (100%)
cl367	0.76	121	1206	ASC1 (58%)	Brain (100%)
cl224	0.76	21	206	No annotation (100%)	hnLN (100%)
cl147	0.76	414	4144	Mid Erythroid (28.2%)	Liver (99.9%)
cl463	0.76	1431	14305	CD8+/CD45RA+ Naive Cytotoxic (1.1%)	Blood (100%)
cl128	0.75	611	6107	TA 2 (39%)	Colon (100%)
cl621	0.75	653	6529	CD4+ Activated Fos-lo (36.6%)	Colon (100%)
cl303	0.75	36	355	SCT (97.5%)	Placenta (100%)
cl31	0.75	51	514	No annotation (100%)	axLN (100%)
cl485	0.75	445	4453	No annotation (100%)	BoneMarrow (100%)
cl52	0.75	5	52	Fibroblasts (86.5%)	Upper airway (100%)
cl39	0.74	103	1034	No annotation (100%)	Omentum Adipose Tissue (100%)
cl473	0.74	124	1237	CD4 (46.6%)	Blood (100%)
cl58	0.74	179	1793	B cell memory (38.8%)	mLN (100%)
cl225	0.74	24	237	No annotation (100%)	hnLN (100%)
cl519	0.74	1064	10640	No annotation (100%)	BoneMarrow (100%)
cl132	0.74	23	229	No annotation (100%)	hnLN (100%)

Table B.14: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 6)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl496	0.74	281	2810	No annotation (100%)	BoneMarrow (100%)
cl285	0.73	129	1290	CD4 Tfh (41.4%)	mLN (100%)
cl623	0.73	122	1224	TA 2 (31.9%)	Colon (100%)
cl434	0.73	867	8667	No annotation (100%)	Prostate (100%)
cl16	0.73	39	389	Endothelium (56.3%)	Lung Parenchyma (100%)
cl593	0.73	4	38	Best4+ Enterocytes (100%)	Colon (100%)
cl148	0.73	15	147	No annotation (100%)	hnLN (100%)
cl54	0.73	122	1217	ODC1 (93.1%)	Brain (100%)
cl513	0.72	659	6588	No annotation (100%)	BoneMarrow (100%)
cl441	0.72	759	7588	CD8+/CD45RA+ Naive Cytotoxic (0.9%)	Blood (100%)
cl248	0.72	27	267	No annotation (100%)	hnLN (100%)
cl173	0.72	13	127	No annotation (100%)	hnLN (100%)
cl454	0.72	1072	10715	CD8+/CD45RA+ Naive Cytotoxic (0.3%)	Blood (100%)
cl598	0.72	22	219	Immature Goblet (77.6%)	Colon (100%)
cl409	0.72	1576	15761	CD4+/CD45RO+ Memory (44.6%)	Blood (100%)
cl315	0.72	46	461	exPFC1 (45.3%)	Brain (100%)
cl555	0.71	506	5061	Immature Goblet (28.8%)	Colon (100%)
cl212	0.71	27	267	DC1 (89.9%)	Decidua (100%)
cl416	0.71	19	192	No annotation (100%)	Prostate (100%)
cl0	0.70	23	231	Endo (f) (34.2%)	Placenta (100%)
cl337	0.70	79	789	exDG (87.6%)	Brain (100%)
cl174	0.70	79	785	No annotation (100%)	axLN (100%)
cl415	0.70	1267	12670	PB Naive CD4 (25.7%)	Blood (100%)
cl205	0.70	781	7807	Nephron_epithelium (6.6%)	Kidney (100%)
cl476	0.70	218	2183	CD8+ Cytotoxic T (83.3%)	Blood (100%)
cl385	0.69	152	1523	No annotation (100%)	Omentum Adipose Tissue (100%)
cl274	0.69	60	600	No annotation (100%)	axLN (100%)
cl427	0.68	149	1487	No annotation (100%)	axLN (100%)
cl53	0.68	481	4814	Granulocytes (2%)	Blood (100%)
cl137	0.68	28	275	No annotation (100%)	axLN (100%)
cl180	0.68	107	1066	No annotation (100%)	axLN (100%)
cl574	0.67	330	3304	Plasma (94.3%)	Colon (100%)
cl185	0.67	155	1554	B cell follicular (53%)	mLN (100%)
cl622	0.67	134	1338	Immature Enterocytes 2 (52.9%)	Colon (100%)
cl546	0.67	119	1191	No annotation (100%)	BoneMarrow (100%)
cl213	0.67	105	1051	No annotation (100%)	axLN (100%)
cl384	0.67	4	43	No annotation (100%)	Omentum Adipose Tissue (100%)
cl28	0.66	170	1703	No annotation (100%)	BoneMarrow (100%)
cl472	0.66	405	4046	CD8+ T cells (20%)	Blood (100%)
cl325	0.66	27	274	No annotation (100%)	Brain_Microglia (100%)
cl184	0.66	31	311	No annotation (100%)	axLN (100%)
cl152	0.66	20	203	Ciliated (100%)	Lung Parenchyma (100%)
cl423	0.66	35	345	No annotation (100%)	axLN (100%)
cl443	0.65	971	9705	Kupffer Cell (69.9%)	Liver (100%)
cl479	0.65	248	2480	Bcell (1.8%)	Blood (100%)
cl365	0.65	74	740	ODC1 (93.9%)	Brain (100%)
cl410	0.65	1714	17137	PB Naive CD4 (0.2%)	Blood (100%)
cl450	0.65	676	6759	CD4+ T Helper (43.9%)	Blood (100%)
cl624	0.65	108	1078	Immature Enterocytes 1 (78.7%)	Colon (100%)
cl218	0.65	54	544	No annotation (100%)	axLN (100%)
cl156	0.65	1652	16515	PB Naive CD4 (0.1%)	Blood (100%)
cl541	0.65	176	1756	No annotation (100%)	BoneMarrow (100%)
cl141	0.65	12	121	No annotation (100%)	Intestine (100%)

Table B.15: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 7)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl221	0.65	1297	12967	PB Naive CD4 (0.3%)	Blood (100%)
cl580	0.64	633	6331	NK cell (6.7%)	Kidney (100%)
cl215	0.64	22	217	No annotation (100%)	hnLN (100%)
cl414	0.64	88	883	Kupffer Cell (29.7%)	Liver (100%)
cl145	0.64	65	646	No annotation (100%)	hnLN (100%)
cl484	0.64	13	134	No annotation (100%)	hnLN (100%)
cl251	0.64	91	906	B cell memory (47.7%)	mLN (100%)
cl176	0.63	42	416	No annotation (100%)	hnLN (100%)
cl33	0.63	15	152	No annotation (100%)	Brain_Microglia (100%)
cl614	0.63	266	2655	CD4+ Memory (41.7%)	Colon (100%)
cl297	0.63	11	113	ILC3 (92%)	Tonsil (100%)
cl124	0.63	392	3915	Plasma (43.6%)	Colon (100%)
cl144	0.62	27	268	No annotation (100%)	hnLN (100%)
cl241	0.62	61	614	No annotation (100%)	axLN (100%)
cl187	0.62	42	418	No annotation (100%)	hnLN (100%)
cl172	0.62	100	1003	No annotation (100%)	axLN (100%)
cl175	0.62	75	748	No annotation (100%)	axLN (100%)
cl420	0.61	134	1337	No annotation (100%)	axLN (100%)
cl273	0.61	12	117	CD4 T central memory (81.2%)	mLN (100%)
cl346	0.60	7	72	Unknown4 (59.7%)	Brain (100%)
cl619	0.59	87	869	CD4+ Activated Fos-hi (48.4%)	Colon (100%)
cl591	0.59	328	3283	Plasma (48.4%)	Colon (100%)
cl529	0.58	39	389	No annotation (100%)	BoneMarrow (100%)
cl231	0.58	17	167	No annotation (100%)	hnLN (100%)
cl597	0.58	24	238	Enterocytes (41.6%)	Colon (100%)
cl550	0.58	113	1128	No annotation (100%)	BoneMarrow (100%)
cl265	0.57	26	262	CD4 T central memory (83.2%)	mLN (100%)
cl368	0.57	23	229	No annotation (100%)	Brain_Microglia (100%)
cl324	0.57	7	73	No annotation (100%)	Brain_Microglia (100%)
cl197	0.56	44	436	No annotation (100%)	hnLN (100%)
cl216	0.56	80	797	No annotation (100%)	axLN (100%)
cl190	0.56	29	289	No annotation (100%)	hnLN (100%)
cl73	0.55	374	3741	B cell IgA plasma (21.1%)	Colon (100%)
cl366	0.55	78	782	exPFC1 (61.8%)	Brain (100%)
cl106	0.55	35	347	CD4 T central memory (73.5%)	mLN (100%)
cl608	0.55	51	512	Macrophages (28.1%)	Colon (100%)
cl584	0.54	79	794	ASC1 (50.5%)	Brain (100%)
cl419	0.54	121	1208	No annotation (100%)	axLN (100%)
cl620	0.54	151	1509	Plasma (67.1%)	Colon (100%)
cl153	0.53	136	1355	No annotation (100%)	axLN (100%)
cl445	0.52	1945	19446	PB Naive CD4 (0.2%)	Blood (100%)
cl625	0.52	91	907	Enterocyte Progenitors (62.7%)	Colon (100%)
cl573	0.52	32	315	ODC1 (89.2%)	Brain (100%)
cl290	0.52	26	262	No annotation (100%)	Brain_Microglia (100%)
cl446	0.52	911	9111	PB Naive CD8 (0.3%)	Blood (100%)
cl162	0.50	14	144	No annotation (100%)	hnLN (100%)
cl200	0.50	3	27	DC1 (100%)	Decidua (100%)
cl226	0.50	28	277	No annotation (100%)	hnLN (100%)
cl302	0.50	86	856	VCT (62.4%)	Placenta (100%)
cl559	0.47	37	365	ODC1 (95.9%)	Brain (100%)
cl425	0.47	57	574	No annotation (100%)	axLN (100%)
cl522	0.47	19	193	No annotation (100%)	BoneMarrow (100%)
cl426	0.47	71	709	No annotation (100%)	axLN (100%)
cl10	0.46	9	95	No annotation (100%)	axLN (100%)
cl349	0.46	18	180	exCA3 (72.8%)	Brain (100%)
cl223	0.46	31	312	No annotation (100%)	hnLN (100%)
cl222	0.45	30	296	No annotation (100%)	hnLN (100%)
cl237	0.45	68	677	B cell memory (34.4%)	mLN (100%)

Table B.16: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 8)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl42	0.44	7	66	No annotation (100%)	Brain_Microglia (100%)
cl320	0.43	44	444	exDG (68.7%)	Brain (100%)
cl596	0.43	38	383	CD4+ Memory (35.5%)	Colon (100%)
cl189	0.43	14	143	No annotation (100%)	hnLN (100%)
cl217	0.42	18	176	No annotation (100%)	hnLN (100%)
cl126	0.40	4	40	Mast cell (92.5%)	Lung Parenchyma (100%)
cl20	0.40	21	208	B cell memory (38.5%)	mLN (100%)
cl253	0.38	21	205	No annotation (100%)	hnLN (100%)
cl85	0.38	20	198	No annotation (100%)	axLN (100%)
cl9	0.38	9	86	No annotation (100%)	hnLN (100%)
cl287	0.37	10	104	No annotation (100%)	Brain_Microglia (100%)
cl411	0.37	34	340	Mid Erythroid (46.8%)	Liver (100%)
cl353	0.37	21	205	exPFC1 (95.1%)	Brain (100%)
cl424	0.37	46	459	No annotation (100%)	axLN (100%)
cl169	0.36	19	194	No annotation (100%)	hnLN (100%)
cl609	0.36	85	848	Enterocyte Progenitors (28.8%)	Colon (100%)
cl364	0.35	14	139	No annotation (100%)	Brain_Microglia (100%)
cl351	0.33	20	201	exPFC1 (92%)	Brain (100%)
cl481	0.33	21	205	CD8+/CD45RA+ Naive Cytotoxic (0.5%)	Blood (100%)
cl594	0.32	43	425	TA 1 (38.1%)	Colon (100%)
cl146	0.32	45	453	No annotation (100%)	hnLN (100%)
cl357	0.32	23	234	exDG (71.8%)	Brain (100%)
cl348	0.32	16	158	ASC1 (19.6%)	Brain (100%)
cl310	0.29	6	58	CD4 T central memory (34.5%)	mLN (100%)
cl245	0.28	119	1186	No annotation (100%)	axLN (100%)
cl395	0.27	17	168	No annotation (100%)	Brain_Microglia (100%)
cl421	0.27	16	161	No annotation (100%)	axLN (100%)
cl129	0.27	12	116	No annotation (100%)	Intestine (100%)
cl234	0.25	14	140	No annotation (100%)	hnLN (100%)
cl354	0.25	21	208	exCA3 (65.9%)	Brain (100%)
cl5	0.25	7	71	Tcm (36.6%)	Skin (100%)
cl181	0.22	8	78	No annotation (100%)	Intestine (100%)
cl565	0.22	33	332	ODC1 (92.2%)	Brain (100%)
cl168	0.20	12	115	No annotation (100%)	Intestine (100%)
cl194	0.19	28	276	No annotation (100%)	hnLN (100%)
cl356	0.17	21	214	exPFC1 (94.9%)	Brain (100%)
cl347	0.15	12	123	exPFC1 (86.2%)	Brain (100%)
cl151	0.15	35	349	No annotation (100%)	hnLN (100%)
cl171	0.14	67	666	No annotation (100%)	axLN (100%)
cl149	0.13	138	1376	No annotation (100%)	axLN (100%)
cl186	0.12	14	138	No annotation (100%)	hnLN (100%)
cl230	0.11	101	1006	CD4 T central memory (42.7%)	mLN (100%)
cl383	0.11	18	179	No annotation (100%)	Omentum Adipose Tissue (100%)
cl352	0.10	20	202	exDG (81.2%)	Brain (100%)
cl135	0.09	4	43	B cell (9.3%)	Intestine (90.7%)
cl394	0.07	17	171	No annotation (100%)	Brain_Microglia (100%)
cl103	0.00	0	3	dS2 (100%)	Decidua (100%)
cl104	0.00	0	3	CD8 T cell (66.7%)	mLN (100%)
cl105	0.00	0	3	B cell follicular (66.7%)	mLN (100%)
cl107	0.00	0	4	CD4 T central memory (50%)	mLN (100%)
cl108	0.00	0	3	dS3 (100%)	Decidua (100%)
cl111	0.00	0	4	VCT (100%)	Placenta (100%)
cl112	0.00	0	3	No annotation (100%)	axLN (100%)
cl114	0.00	0	3	Neutrophils (100%)	Lung Parenchyma (100%)
cl116	0.00	0	4	VCT (100%)	Placenta (100%)
cl117	0.00	0	3	B cell memory (66.7%)	mLN (100%)
cl119	0.00	0	3	Endothelium (100%)	Lung Parenchyma (100%)

Table B.17: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 9)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl120	0.00	0	3	Endo (m) (100%)	Decidua (100%)
cl121	0.00	0	3	Type 2 (66.7%)	Lung Parenchyma (100%)
cl123	0.00	0	5	dNK p (60%)	Decidua (100%)
cl125	0.00	1	15	B cell (100%)	Lung Parenchyma (100%)
cl130	0.00	0	5	No annotation (100%)	axLN (100%)
cl136	0.00	7	75	Treg NL-like (41.3%)	mLN (100%)
cl138	0.00	7	73	No annotation (100%)	Intestine (100%)
cl139	0.00	0	3	EVT (100%)	Placenta (100%)
cl140	0.00	0	4	No annotation (100%)	Brain_Microglia (100%)
cl142	0.00	0	4	Macrophages (100%)	Lung Parenchyma (100%)
cl150	0.00	10	103	No annotation (100%)	Intestine (100%)
cl154	0.00	0	4	CD4 T central memory (75%)	mLN (100%)
cl157	0.00	0	5	fFB1 (100%)	Placenta (100%)
cl158	0.00	1	14	B cell (100%)	Lung Parenchyma (100%)
cl159	0.00	0	4	Endo L (100%)	Decidua (100%)
cl161	0.00	11	115	No annotation (100%)	Intestine (100%)
cl163	0.00	3	31	B cell memory (83.9%)	mLN (100%)
cl164	0.00	9	93	No annotation (100%)	Intestine (100%)
cl165	0.00	7	67	No annotation (100%)	Intestine (100%)
cl166	0.00	8	81	No annotation (100%)	Intestine (100%)
cl167	0.00	6	57	No annotation (100%)	Intestine (100%)
cl17	0.00	0	3	No annotation (100%)	Brain_Microglia (100%)
cl170	0.00	7	69	No annotation (100%)	Intestine (100%)
cl177	0.00	9	91	No annotation (100%)	Intestine (100%)
cl182	0.00	1	13	B cell memory (69.2%)	mLN (100%)
cl188	0.00	7	68	No annotation (100%)	Intestine (100%)
cl191	0.00	4	37	No annotation (100%)	Intestine (100%)
cl193	0.00	4	42	No annotation (100%)	Intestine (100%)
cl195	0.00	7	67	No annotation (100%)	Intestine (100%)
cl196	0.00	6	64	No annotation (100%)	Intestine (100%)
cl199	0.00	1	7	Secretory (100%)	Upper airway (100%)
cl201	0.00	1	9	Plasma (100%)	Decidua (100%)
cl202	0.00	7	68	No annotation (100%)	Intestine (100%)
cl204	0.00	8	82	No annotation (100%)	Intestine (100%)
cl207	0.00	5	53	No annotation (100%)	Intestine (100%)
cl209	0.00	1	10	B cell (100%)	Lung Parenchyma (100%)
cl21	0.00	2	22	No annotation (100%)	Intestine (100%)
cl211	0.00	0	4	No annotation (100%)	Brain_Microglia (100%)
cl228	0.00	11	114	No annotation (100%)	hnLN (100%)
cl235	0.00	1	9	B cell follicular (33.3%)	mLN (100%)
cl236	0.00	1	8	Type 1 (100%)	Lung Parenchyma (100%)
cl239	0.00	0	3	No annotation (100%)	Brain_Microglia (100%)
cl244	0.00	1	6	Granulocytes (100%)	Decidua (100%)
cl249	0.00	18	182	No annotation (100%)	axLN (100%)
cl257	0.00	1	15	VCT (100%)	Placenta (100%)
cl288	0.00	8	79	No annotation (100%)	Brain_Microglia (100%)
cl289	0.00	5	52	No annotation (100%)	Brain_Microglia (100%)
cl29	0.00	0	4	B cell follicular (75%)	mLN (100%)
cl291	0.00	1	7	dS1 (100%)	Decidua (100%)
cl292	0.00	2	24	No annotation (100%)	Brain_Microglia (100%)
cl294	0.00	1	14	No annotation (100%)	Brain_Microglia (100%)
cl300	0.00	26	259	Ciliated (97.7%)	Upper airway (100%)
cl305	0.00	1	6	No annotation (100%)	Omentum Adipose Tissue (100%)
cl32	0.00	0	3	Endo (m) (100%)	Placenta (100%)
cl328	0.00	12	121	No annotation (100%)	Brain_Microglia (100%)
cl330	0.00	8	78	No annotation (100%)	Brain_Microglia (100%)
cl331	0.00	11	109	No annotation (100%)	Brain_Microglia (100%)
cl333	0.00	11	107	No annotation (100%)	Brain_Microglia (100%)

Table B.18: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 10)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl334	0.00	12	119	CD4 Tfh (73.1%)	mLN (100%)
cl342	0.00	0	5	exCA3 (80%)	Brain (100%)
cl344	0.00	2	21	GABA2 (57.1%)	Brain (100%)
cl363	0.00	6	57	No annotation (100%)	Brain_Microglia (100%)
cl381	0.00	1	7	Ciliated (100%)	Lung Parenchyma (100%)
cl393	0.00	3	35	No annotation (100%)	Brain_Microglia (100%)
cl396	0.00	2	22	No annotation (100%)	Brain_Microglia (100%)
cl397	0.00	4	41	No annotation (100%)	Brain_Microglia (100%)
cl398	0.00	4	43	No annotation (100%)	Brain_Microglia (100%)
cl405	0.00	3	26	Kupffer Cell (69.2%)	Liver (100%)
cl406	0.00	0	5	ILC precursor (100%)	Liver (100%)
cl407	0.00	1	8	Kupffer Cell (62.5%)	Liver (100%)
cl436	0.00	0	3	CD8+/CD45RA+ Naive Cytotoxic (100%)	Blood (100%)
cl437	0.00	0	3	CD4+/CD25 T Reg (100%)	Blood (100%)
cl461	0.00	0	3	CD4+/CD25 T Reg (100%)	Blood (100%)
cl462	0.00	2	20	CD4+/CD45RA+/CD25- Naive T (55%)	Blood (100%)
cl466	0.00	0	3	CD4+/CD25 T Reg (100%)	Blood (100%)
cl467	0.00	0	4	CD56+ NK (50%)	Blood (100%)
cl468	0.00	1	7	No annotation (100%)	Blood (100%)
cl469	0.00	1	10	No annotation (100%)	Blood (100%)
cl471	0.00	2	24	No annotation (100%)	Blood (100%)
cl482	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl487	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl489	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl499	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl521	0.00	2	19	No annotation (100%)	BoneMarrow (100%)
cl525	0.00	13	129	No annotation (100%)	BoneMarrow (100%)
cl526	0.00	4	43	No annotation (100%)	BoneMarrow (100%)
cl527	0.00	3	29	No annotation (100%)	BoneMarrow (100%)
cl528	0.00	1	8	No annotation (100%)	BoneMarrow (100%)
cl530	0.00	0	5	No annotation (100%)	BoneMarrow (100%)
cl531	0.00	0	5	No annotation (100%)	BoneMarrow (100%)
cl532	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl533	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl535	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl539	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl544	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl545	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl557	0.00	3	29	No annotation (100%)	Brain_Microglia (100%)
cl558	0.00	13	131	No annotation (100%)	Brain_Microglia (100%)
cl560	0.00	12	118	No annotation (100%)	Brain_Microglia (100%)
cl561	0.00	10	101	No annotation (100%)	Brain_Microglia (100%)
cl562	0.00	4	39	No annotation (100%)	Brain_Microglia (100%)
cl563	0.00	3	30	No annotation (100%)	Brain_Microglia (100%)
cl564	0.00	3	28	No annotation (100%)	Brain_Microglia (100%)
cl566	0.00	0	3	No annotation (100%)	Brain_Microglia (100%)
cl572	0.00	13	133	No annotation (100%)	Brain_Microglia (100%)
cl578	0.00	1	12	Endothelium; Ascending_vasa_recta; VCAM1- (8.3%)	Kidney (100%)
cl579	0.00	15	154	No annotation (100%)	hnLN (100%)
cl583	0.00	0	3	dS2 (100%)	Decidua (100%)
cl585	0.00	0	3	No annotation (100%)	BoneMarrow (100%)
cl587	0.00	1	12	ODC1 (100%)	Brain (100%)
cl588	0.00	2	19	ODC1 (94.7%)	Brain (100%)
cl589	0.00	9	88	ASC1 (48.9%)	Brain (100%)
cl599	0.00	10	101	Immature Goblet (81.2%)	Colon (100%)

Table B.19: F1 scores and class sizes for *CellTypist* trained on the human collection with integrated cluster labels. Labels are derived from the *CellTypist* pipeline, as described in Chapter 3. "Major Cell Type" refers to the most represented cell type, not counting cell without annotation, unless none have it. (continued 11)

Cluster	F1 Score	Test Support	Total Cells	Major Cell Type	Major Tissue
cl6	0.00	3	34	T cell (41.2%)	Upper airway (100%)
cl600	0.00	2	21	TA 2 (85.7%)	Colon (100%)
cl602	0.00	2	20	Immature Enterocytes 1 (35%)	Colon (100%)
cl603	0.00	1	8	Cycling TA (50%)	Colon (100%)
cl604	0.00	0	4	Enterocytes (100%)	Colon (100%)
cl605	0.00	0	3	Immature Goblet (100%)	Colon (100%)
cl607	0.00	0	3	Macrophages (66.7%)	Colon (100%)
cl61	0.00	4	39	No annotation (100%)	Intestine (100%)
cl65	0.00	0	3	Basal (100%)	Upper airway (100%)
cl66	0.00	0	4	No annotation (100%)	Intestine (100%)
cl72	0.00	2	23	B cell (87%)	Upper airway (100%)
cl78	0.00	0	3	dM2 (100%)	Decidua (100%)
cl81	0.00	1	6	dM2 (83.3%)	Decidua (100%)
cl82	0.00	0	3	dT CD4 (33.3%)	Decidua (100%)
cl84	0.00	0	3	Neutrophils (100%)	Lung Parenchyma (100%)
cl86	0.00	0	3	Basal (100%)	Upper airway (100%)
cl90	0.00	0	3	dM2 (100%)	Decidua (100%)
cl91	0.00	0	3	dM2 (66.7%)	Decidua (100%)
cl92	0.00	0	3	VCT (100%)	Decidua (100%)
cl94	0.00	0	3	CD4 Tfh (66.7%)	mLN (100%)
cl95	0.00	0	3	dNK1 (100%)	Decidua (100%)
cl98	0.00	0	5	dM3 (60%)	Placenta (100%)
cl99	0.00	0	3	CD4 Tfh (33.3%)	mLN (100%)

Table B.20: Top genes in the largest merged clusters of each *CellTypist* model

Model	Cluster	Top Genes
thr1 = 0.99, thr2 = 0.8	cl87	S100A4, FOS, KLRB1, DUSP1, NFKBIA, KLF6, LTB, CXCR4, ANXA1, SRGN
	cl147	HBG2, EEF1A1, RNASE1, HMOX1, RPL39,AL138963.3, AC026803.1, H3F3A, EGFL7, FP671120.4
	cl102	C7, DCN, DLK1, IGF2, COL3A1, COL1A1, HSPA1A, TSHZ2, HSPA1B, MMP2
	cl155	IGLL1, VPREB1, HIST1H4C, HMGB2, H3F3A, PTTG1, IL7R, CD24, SMC4, HMGA1
	cl160	MT-RNR2, MT-TT, MT-TG, SNORA31, MT-RNR1, MTCO1P40, MT-TK, EEF1A1P5, MT2A, Y_RNA
thr1 = 0.4, thr2 = 0.99	cl263	RPL10P9, RPS3A, DONSON, RPL9, RPS10, AL031280.1, SELENOM, RPS26, DPY30, RPL7
	cl530	PPBP, MT-RNR1, GNG11, HIST1H2AC, MIR1244-2, NCOA4, GPX1, PF4, OAZ1, CAVIN2
	cl215	GLRX, REXO2, CPVL, GYPB, HIST1H4C, FAM178B, HEMGN, RGS16, TUBA3C, GIHCG
	cl234	GNLY, CD52, NKG7, GZMH, CD3D, CD3G, IL32, TRGC2, TRAC, TRBC1
	cl233	IGLC2, IGLC3, HLA-DRA, CD74, AL365357.1, CD52, MIR1244-2, MTATP6P1, HLA-DQB1, AC005912.1
thr1 = 0.25, thr2 = 0.25	cl114	FN1, TPT1, MARCO, RPL10, SARAF, EEF1A1, PS3A, TIMP3, RPS29, AL365357.1
	cl72	CCL3L1, AL450405.1, RPL41, KLRF1, IGHA1, DUSP4, GZMK, CCL4L1, TYROBP, CCN1
	cl102	MTND1P23, RPS26, JUNB, AL450405.1, MTCO1P12, RPS4Y1, ACTB, C20orf204, LTB, MIR1244-2
	cl10	PLP1, LINC01116, SELE, HMOX1, IGFBP5, CXCL12, MTRNR2L8, TFPI2, HBG1, APOE
	cl23	AL450405.1, HLA-DRA, AC027290.2, RPL26, CD74, RPL39, H3F3A, RPS26, LINC01781, HLA-DRB6
thr1 = 0.1, thr2 = 0.1	cl10	AMH, DHRS2, ADAMDEC1, SELE, CRHBP, AL450405.1, INS, POSTN, TMEM88, GZMK
	cl1	FAM178B, PNMT, GAL, CCL3L1, SFTPB, GCG, RAB38, KLF1, HLA-DRB6, CCL5
	cl2	WFDC1, PHGR1, IGFBP3, PAGE4, BAMBI, MARCO, IGSF6, SERPINB3, FRZB, HAPLN1
	cl20	MTND1P23, AL450405.1, NHSL2, ZNF90, JUNB, CPA5, MTCO1P12, AL513365.1, RPL9P9, RP11-138A9.2
	cl57	AL365226.1, MTRNR2L12, XAGE2, ANAPC4, AC068134.2, IL24, RETREG1, C3, CSF1R, EMX1

Appendix C

Additional information to Chapter 4

This Appendix contains supplementary figures for Chapter 4.

C.1 Supplementary Figures

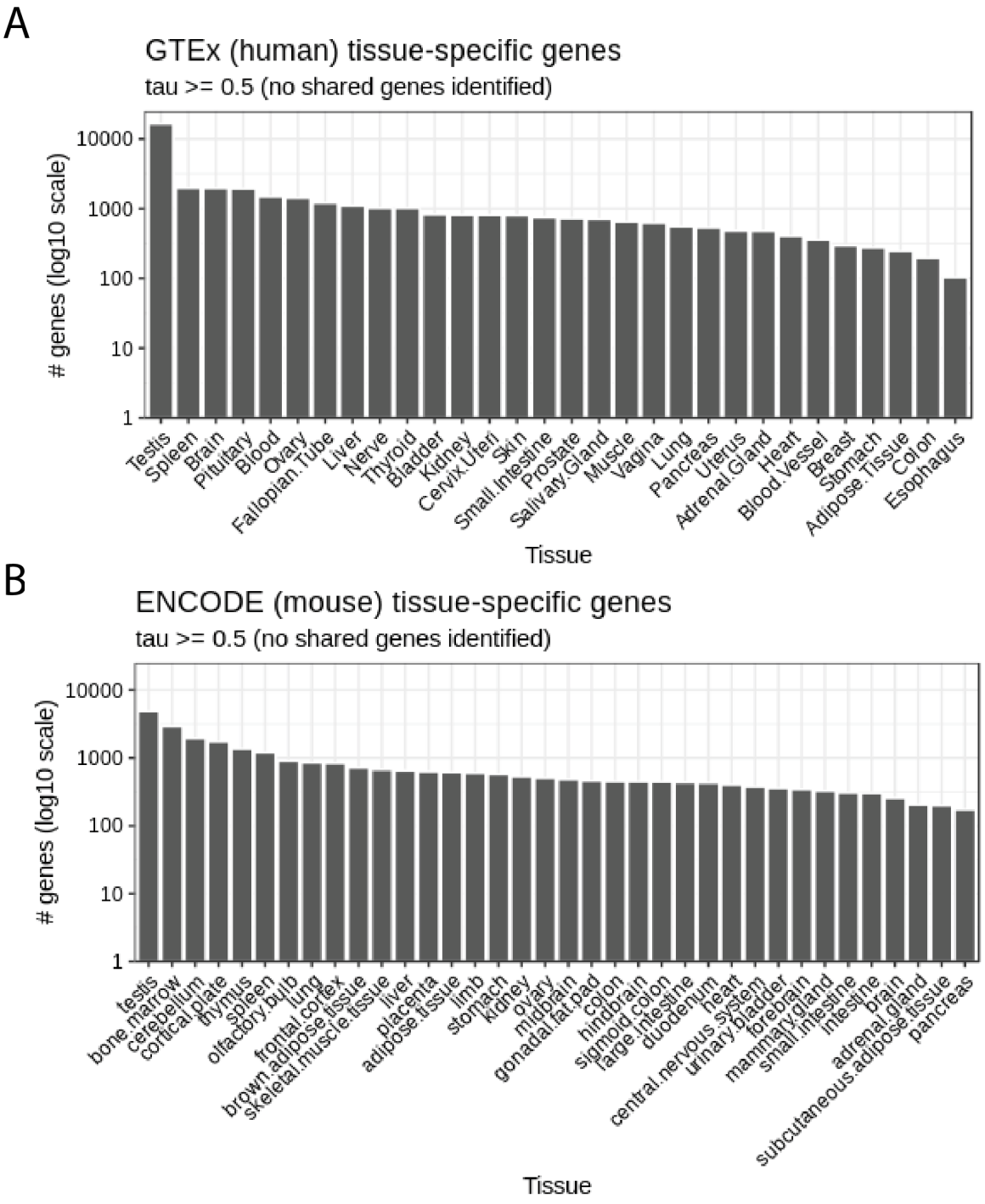


Fig. C.1: **Number of tissue-specific genes determined per tissue for human (A) and mouse (B))**
Tissue specific genes were determined by calculating tau (see Section 4.4.2) and keeping only those with a value greater than 0.5. No genes shared between tissues were found.

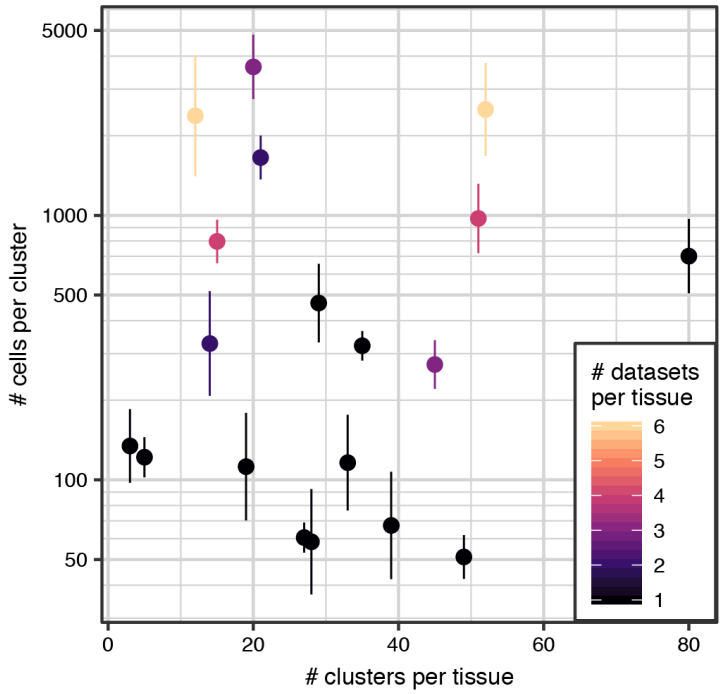


Fig. C.2: **Relating number of per-tissue clusters and number of cells (Related to Figure 3.7A)**
Scatter plot showing the variation of number of clusters per tissue with the number of cells, as well as number of datasets collected for each tissue (colour).

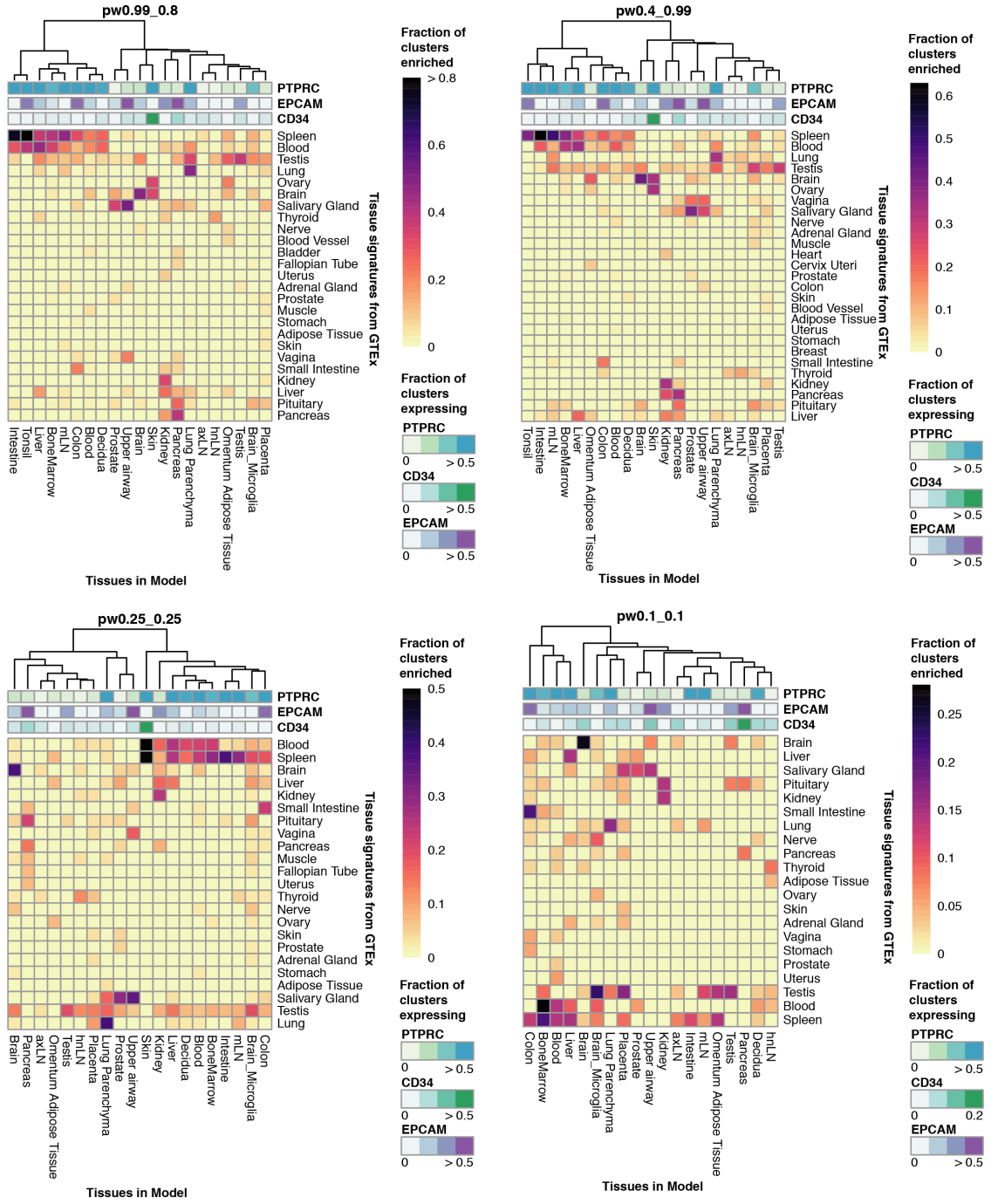


Fig. C.3: **Enrichment of tissue gene modules in other *CellTypist* models (Related to Figure 4.3)**
Heatmaps showing the fraction of clusters in each tissue (x-axis) with an enrichment for tissue-specific gene programmes (y-axis) determined from GTEx data. Each heatmap represents a different set of clusters per tissue, resulting from using different parameters in the *CellTypist* pipeline. Plot for thr1 = 0.99, thr2 = 0.8 is identical to Figure 4.3B.

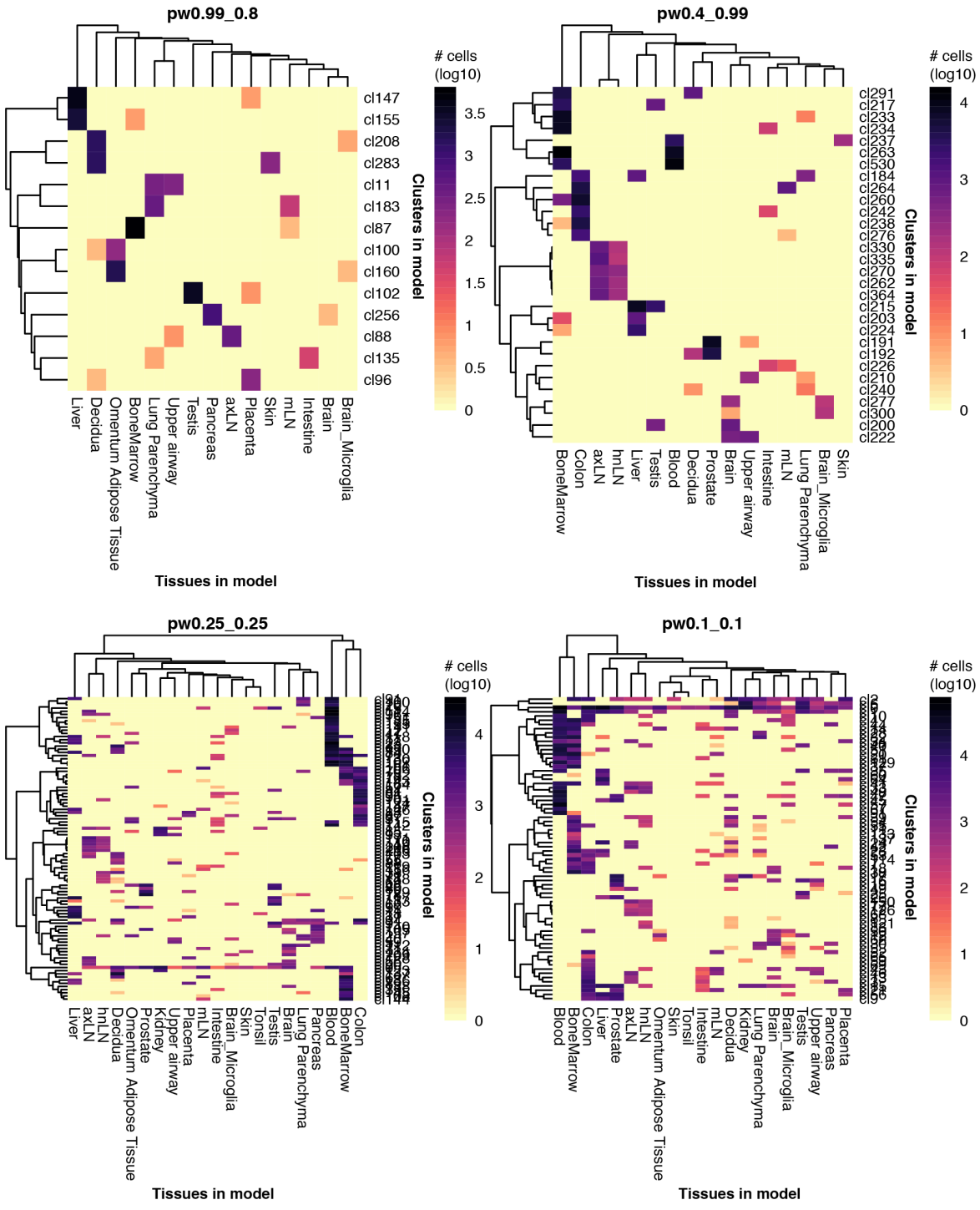


Fig. C.4: **Clusters merged across tissues in the different models (Related to Figure 4.3)**
Heatmaps showing the number of cells contributed by each tissue into cross-tissue clusters for each model.

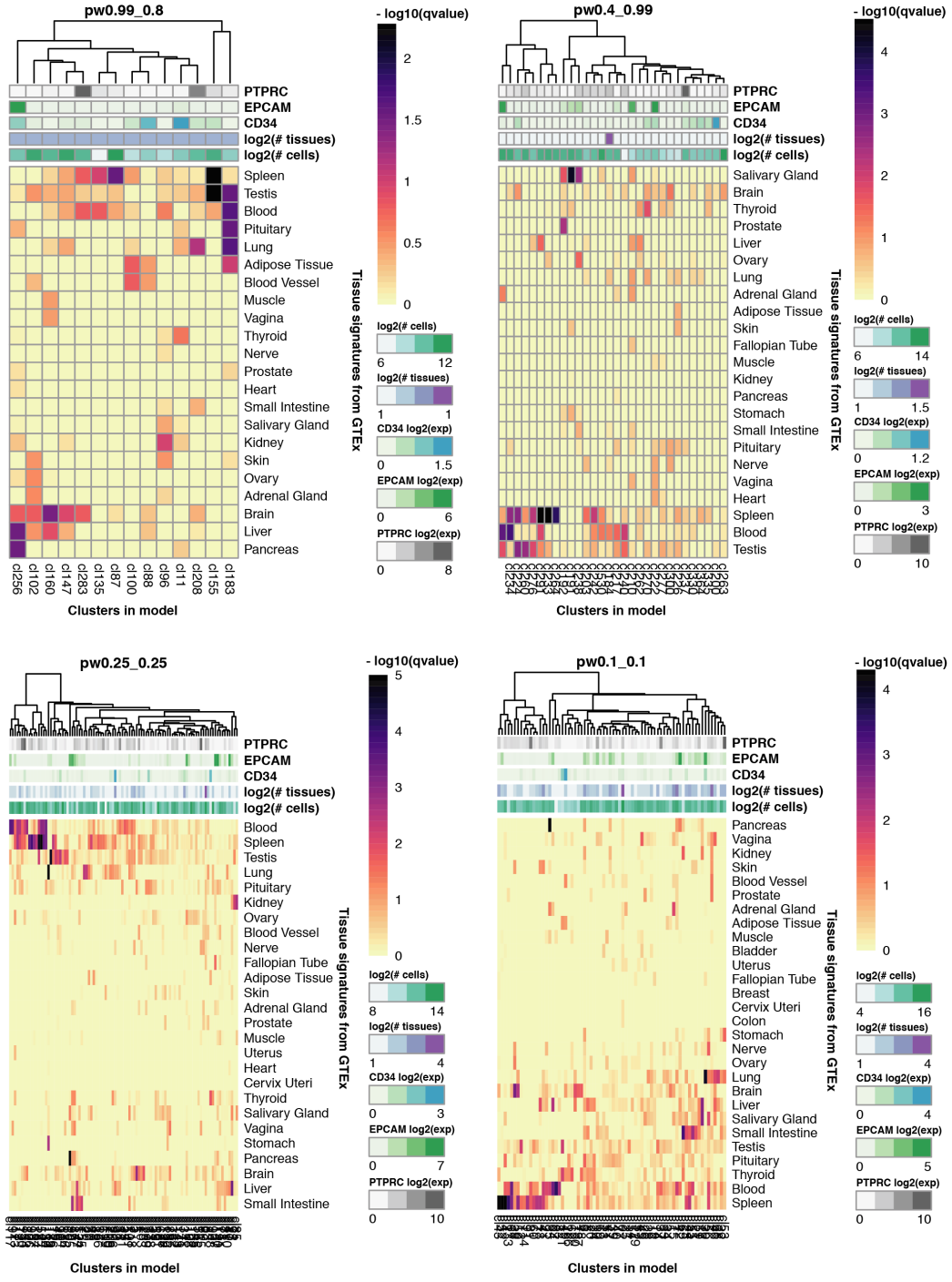


Fig. C.5: **Enrichment of tissue gene modules in merged clusters of different *CellTypist* models (Related to Figure 4.3)**
Heatmaps showing the -log10(q-value) of each merged cluster (x-axis) for the enrichment of tissue-specific gene programmes (y-axis) in their top 500 genes output by the model. Each heatmap represents a different set of merged clusters, resulting from using different parameters in the *CellTypist* pipeline.

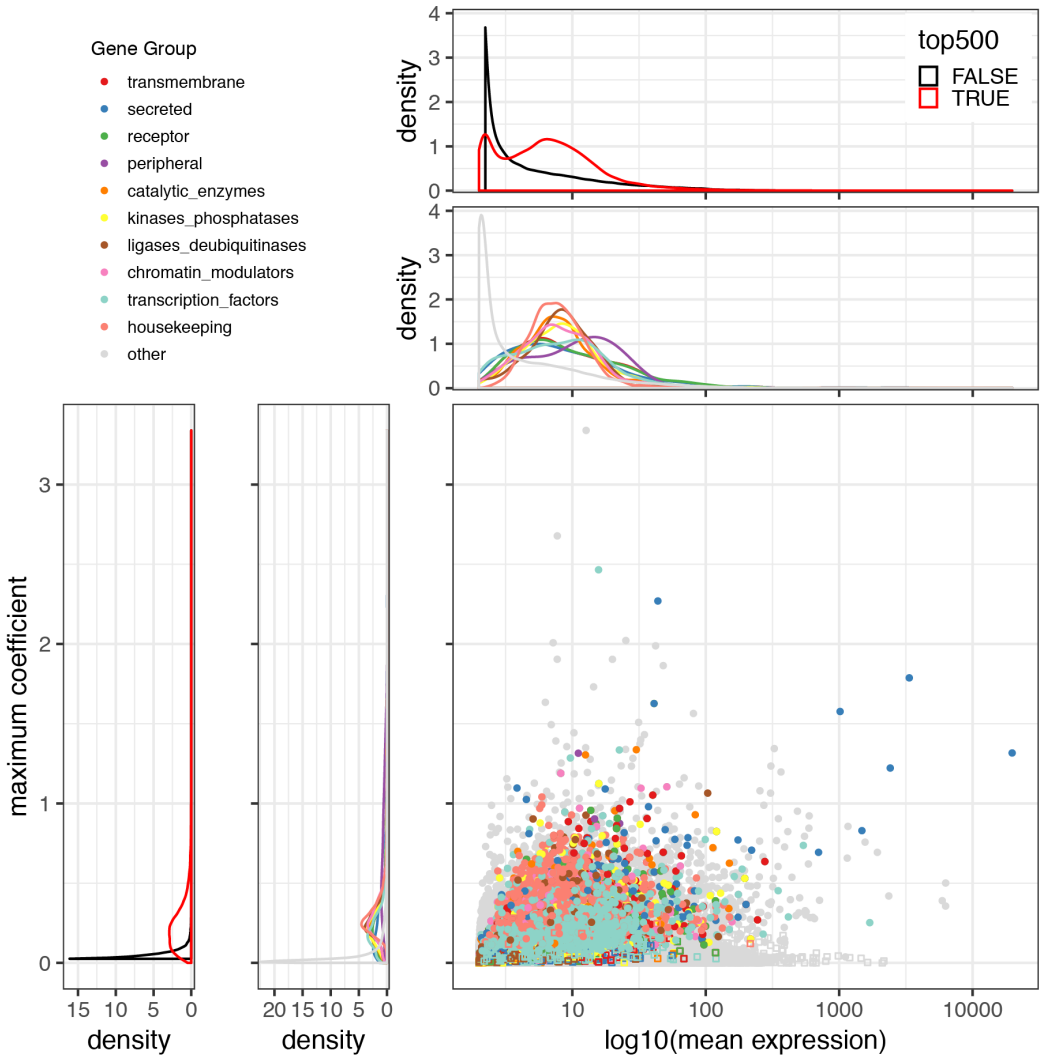


Fig. C.6: **Correlation between gene expression and importance in the human *CellTypist* model (Related to Figure 4.4)**
Scatterplot shows the relationship between mean expression across all cells and the maximum coefficient for each gene across all labels. Density plots show distribution of gene groups, and distribution of genes included in the top 500 coefficients of any label, along the mean expression (top) or maximum coefficient (left) range. Spearman correlation coefficient = 0.56, p-value < 0.01.

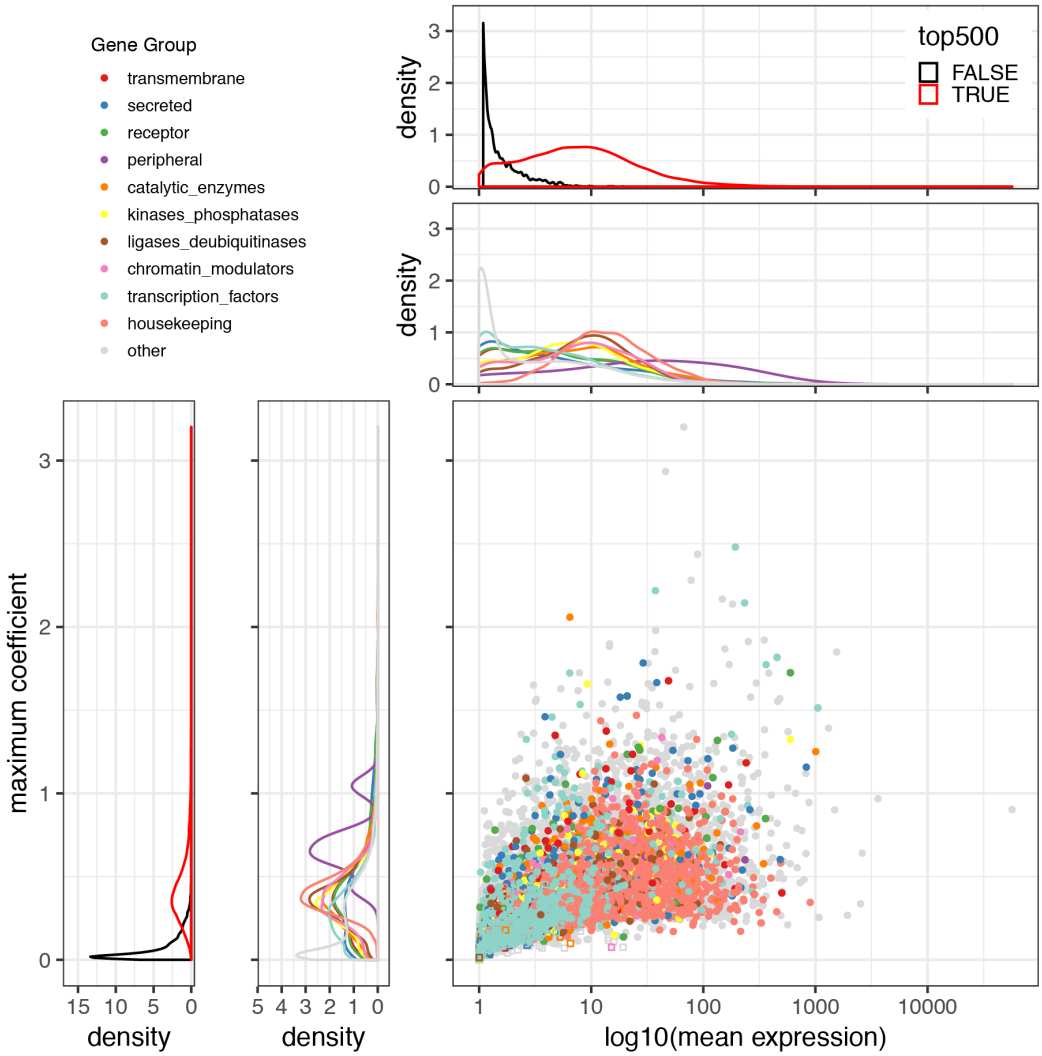


Fig. C.7: **Correlation between gene expression and importance in the *Tabula Muris CellTypist* model (Related to Figure 4.5)**
Scatterplot shows the relationship between mean expression across all cells and the maximum coefficient for each gene across all labels. Density plots show distribution of gene groups, and distribution of genes included in the top 500 coefficients of any label, along the mean expression (top) or maximum coefficient (left) range. Spearman correlation coefficient = 0.86, p-value < 0.01.

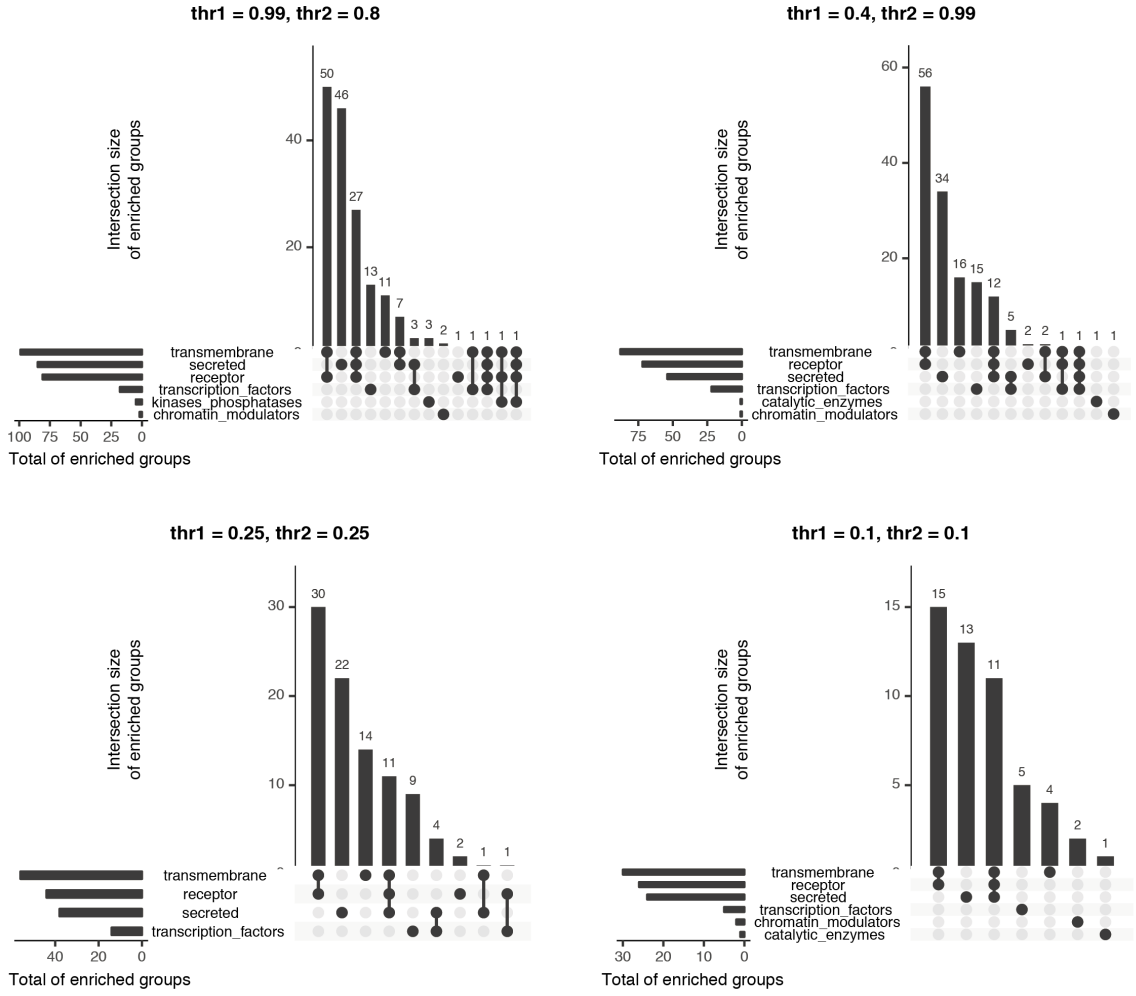


Fig. C.8: **Gene upset plots of different *CellTypist* models (Related to Figure 4.4)** Upset plots counting the number of clusters enriched for a specific group of genes in each model. The gene groups tested were "transcription factors", "transmembrane", "secreted", "receptors", "membrane peripheral proteins", "kinases and phosphatases", "chromatin modulators", "catalytic enzymes", "housekeeping genes". Only the terms enriched in at least one cluster were shown. The plot for thr1 = 0.99, thr2 = 0.8 is identical to Figure 4.4B.

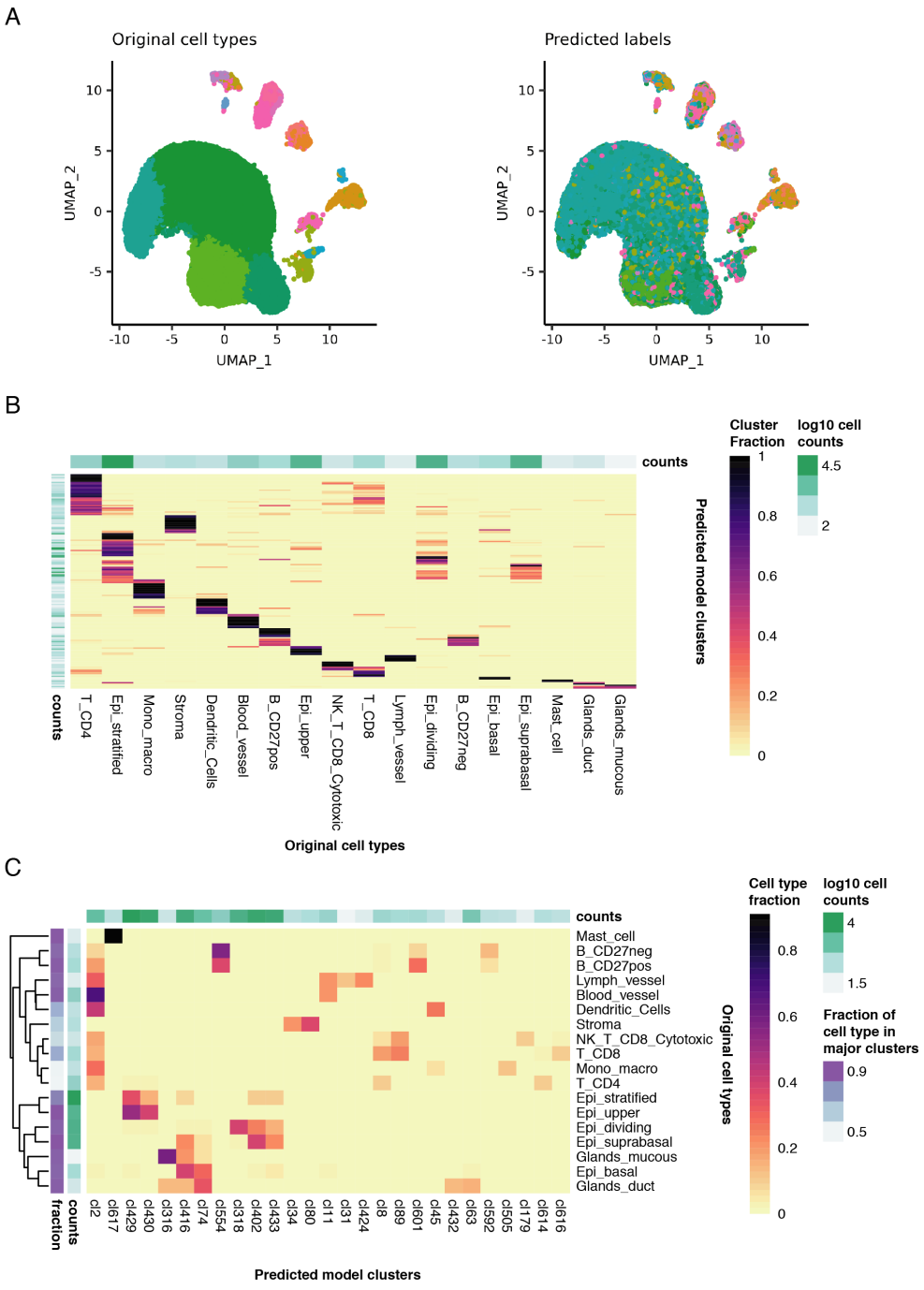


Fig. C.9: ***CellTypist* predictions for oesophagus data from (Madissoon et al., 2019) (Related to Figure 4.1)**
(A) UMAP projections coloured by the original cell type annotations (left) and those predicted by *CellTypist* (right) using thr1 = 0.99 and thr2 = 0.8. **(B)** Proportion of clusters (rows) matching each annotated cell type (columns). **(C)** Proportion of annotated cell types (rows) included in each cluster (columns). Only clusters including at least 10% of a given cell type were included.

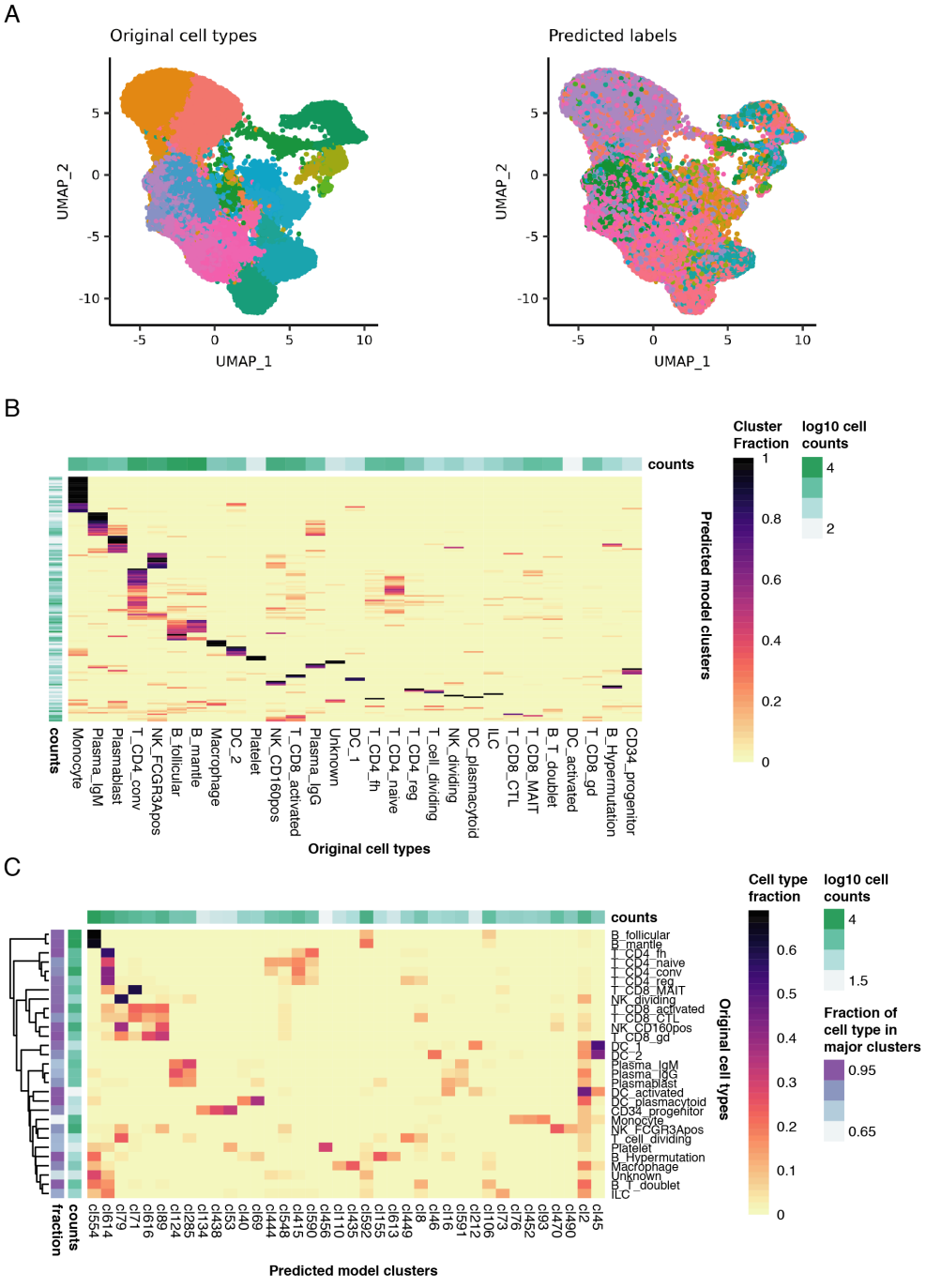


Fig. C.10: ***CellTypist* predictions for spleen data from (Madissoon et al., 2019) (Related to Figure 4.1)**
(A) UMAP projections coloured by the original cell type annotations (left) and those predicted by *CellTypist* (right) using thr1 = 0.99 and thr2 = 0.8. **(B)** Proportion of clusters (rows) matching each annotated cell type (columns). **(C)** Proportion of annotated cell types (rows) included in each cluster (columns). Only clusters including at least 10% of a given cell type were included.

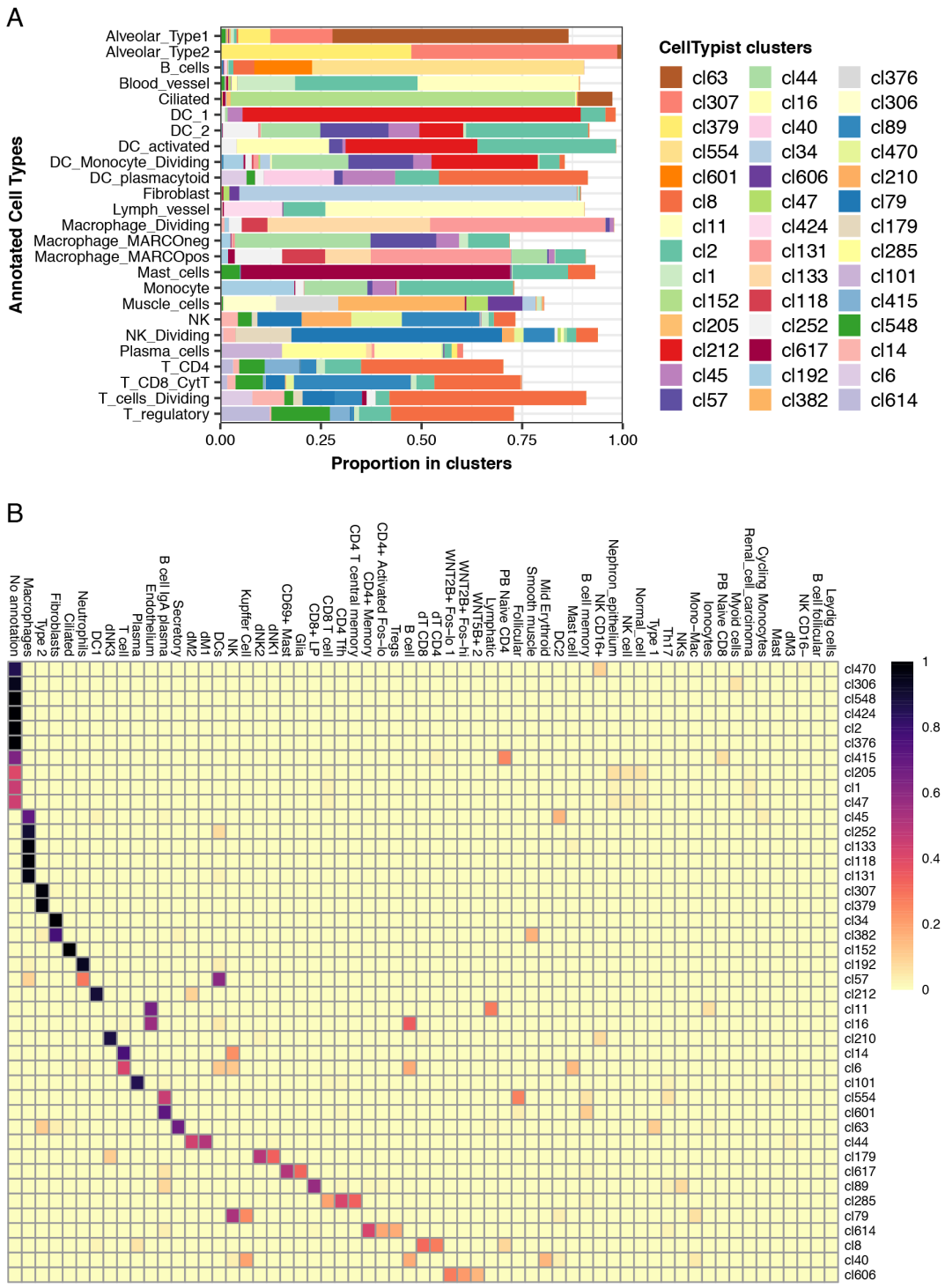


Fig. C.11: **Matching *CellTypist* predictions in lung with annotations in the data collection (Related to Figure 4.1)**
(A) *CellTypist* clusters (thr1 = 0.99, thr2 = 0.8) matched to each original cell type annotation. Only the top 3 clusters per cell type were selected. **(B)** Proportion of cell type annotations (columns) represented in the *CellTypist* clusters matched to lung.

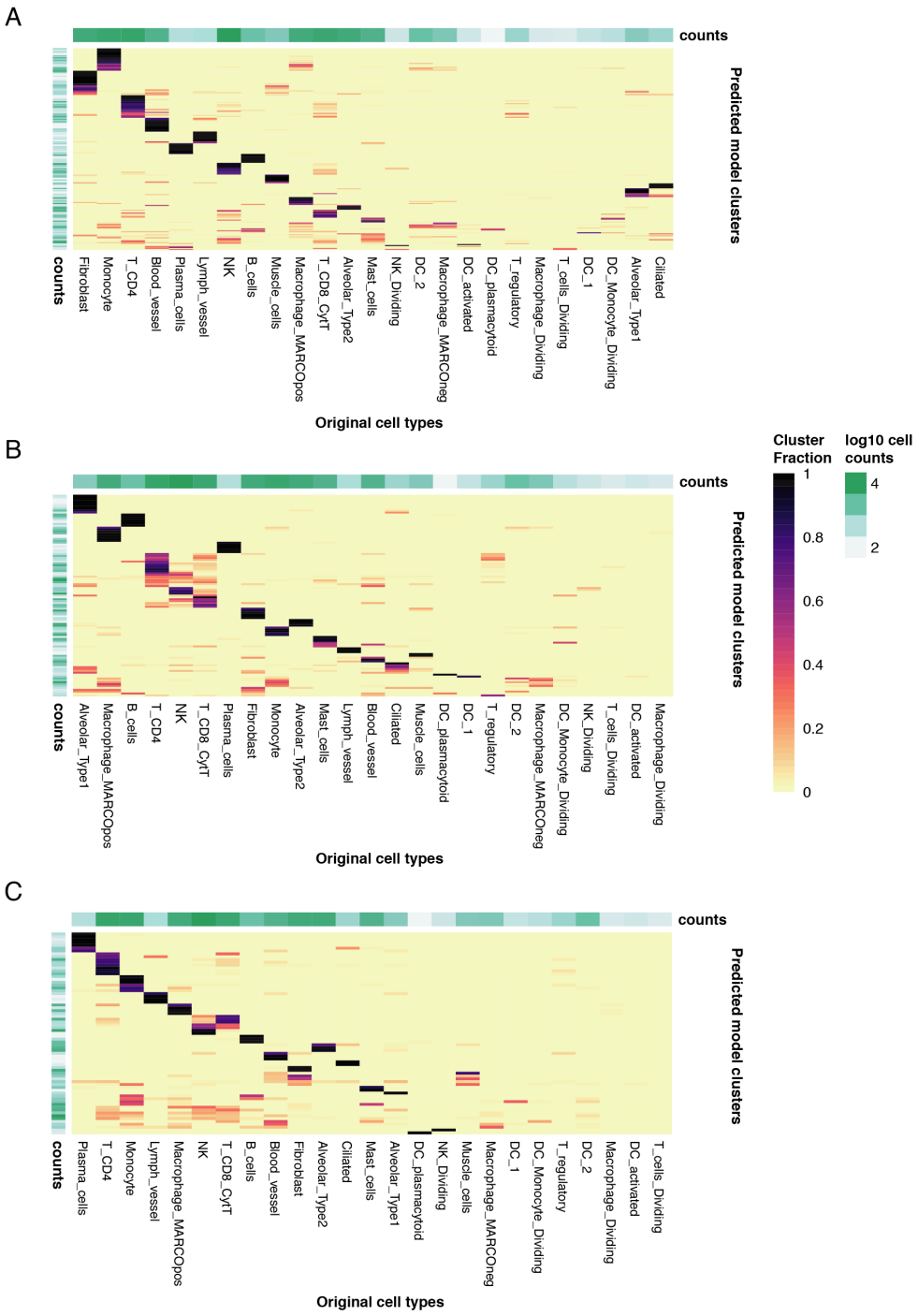


Fig. C.12: **Clusters matching lung annotated cell types in other *CellTypist* models (Related to Figure 4.1B)**
Proportion of clusters (rows) matching each annotated cell type (columns) in the models thr1 = 0.4, thr2 = 0.99 **(A)**, thr1 = 0.25, thr2 = 0.25 **(A)**, and thr1 = 0.1, thr2 = 0.1 **(C)**.

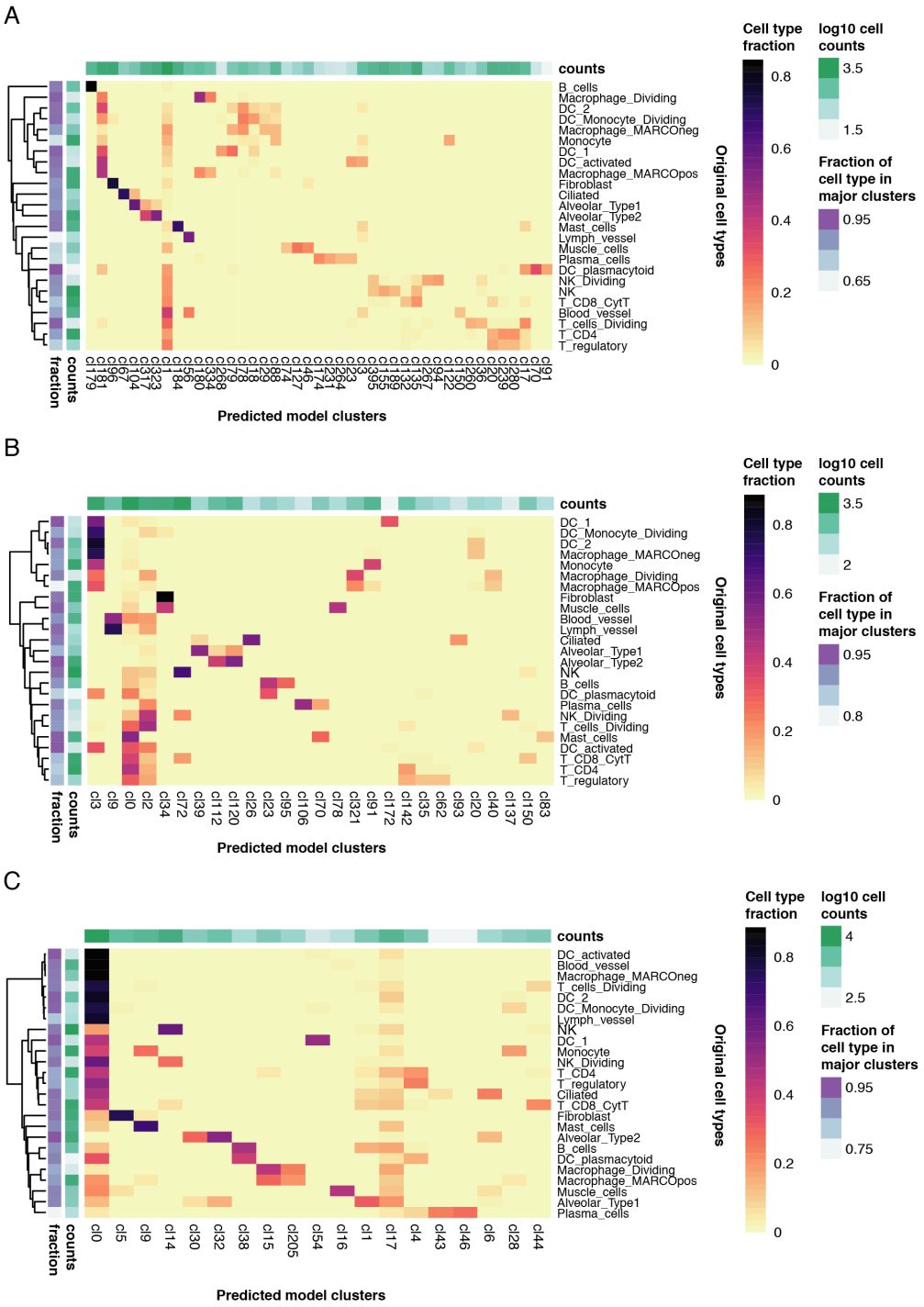


Fig. C.13: **Lung annotated cell types matching clusters in other *CellTypist* models (Related to Figure 4.1C)**
Proportion of annotated cell types (rows) included in each cluster (columns) in the models thr1 = 0.4, thr2 = 0.99 **(A)**, thr1 = 0.25, thr2 = 0.25 **(A)**, and thr1 = 0.1, thr2 = 0.1 **(C)**. Only clusters including at least 10% of a given cell type were included.

C.2 Supplementary Tables

Table C.1: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters

Cluster	Tissue	Cell types
cl430	Lung	Alveolar_Type1
cl430	Oesophagus	Epi_upper,Epi_stratified
cl433	Lung	Alveolar_Type2
cl433	Spleen	Plasmablast,DC_1,Monocyte,NK_dividing,Plasma_IgG
cl429	Lung	Alveolar_Type1
cl39	Lung	T_CD4,T_cells_Dividing,T_regulatory
cl39	Spleen	T_CD4_fh,T_CD4_conv,T_CD4_reg,T_CD8_MAIT,T_CD4_naive
cl39	Oesophagus	T_CD8,T_CD4,NK_T_CD8_Cytotoxic,Mast_cell,Lymph_vessel
cl402	Lung	Alveolar_Type2,Alveolar_Type1
cl402	Oesophagus	Epi_suprabasal
cl318	Lung	T_cells_Dividing,NK_Dividing,DC_Monocyte_Dividing,Macrophage_Dividing,Alveolar_Type1
cl318	Spleen	NK_dividing,T_cell_dividing,B_Hypermutation,Plasmablast,CD34_progenitor
cl318	Oesophagus	Epi_dividing
cl416	Lung	Alveolar_Type1,Alveolar_Type2,Ciliated,Lymph_vessel
cl416	Oesophagus	Glands_mucous,Epi_basal,Glands_duct,Epi_suprabasal
cl263	Lung	Alveolar_Type1,Alveolar_Type2,Ciliated
cl263	Oesophagus	Epi_stratified,Epi_basal,Glands_duct
cl2	Lung	DC_2,DC_activated,Lymph_vessel,DC_Monocyte_Dividing,DC_1
cl2	Spleen	DC_1,DC_activated,DC_2,DC_plasmacytoid,T_CD8_gd
cl2	Oesophagus	Blood_vessel,NK_T_CD8_Cytotoxic,Mast_cell,T_CD8,Dendritic_Cells
cl1	Lung	Blood_vessel
cl1	Oesophagus	Blood_vessel,Mast_cell,Dendritic_Cells,Stroma,NK_T_CD8_Cytotoxic
cl548	Spleen	T_CD8_CTL,T_CD8_MAIT,T_CD4_conv
cl548	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic
cl80	Lung	Fibroblast,Muscle_cells,Lymph_vessel,Blood_vessel
cl80	Spleen	T_CD8_MAIT,T_CD8_CTL
cl80	Oesophagus	Stroma,Epi_basal,Lymph_vessel,Glands_duct,Epi_suprabasal
cl6	Lung	T_cells_Dividing,DC_Monocyte_Dividing,DC_1,DC_activated,DC_plasmacytoid
cl6	Spleen	T_cell_dividing,DC_plasmacytoid,T_CD8_CTL,B_Hypermutation,NK_dividing
cl6	Oesophagus	Dendritic_Cells,NK_T_CD8_Cytotoxic,T_CD8,T_CD4,Mast_cell
cl614	Lung	T_CD4,T_regulatory,T_cells_Dividing
cl614	Spleen	T_CD4_fh,T_CD4_reg,T_CD4_conv,T_CD4_naive
cl614	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic
cl262	Lung	Alveolar_Type1,DC_1,DC_2,DC_activated
cl262	Oesophagus	Epi_stratified
cl63	Lung	Alveolar_Type1,Alveolar_Type2,Ciliated,Blood_vessel,Lymph_vessel
cl63	Spleen	DC_activated,DC_1,CD34_progenitor,DC_2,T_CD8_MAIT
cl63	Oesophagus	Glands_duct,Epi_basal,Glands_mucous,Epi_suprabasal,Lymph_vessel
cl513	Lung	T_CD4,T_cells_Dividing,T_regulatory,Mast_cells
cl513	Spleen	T_CD4_reg,T_CD8_MAIT,T_CD4_conv,T_CD4_fh,T_cell_dividing
cl513	Oesophagus	Mast_cell,T_CD4,NK_T_CD8_Cytotoxic,T_CD8
cl89	Lung	NK_Dividing,T_CD8_CytT,DC_plasmacytoid,DC_activated,NK
cl89	Spleen	T_CD8_activated,T_CD8_gd,T_CD8_MAIT,NK_CD160pos,T_CD8_CTL
cl89	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8,T_CD4,B_CD27pos,Mast_cell
cl377	Lung	Alveolar_Type2,Alveolar_Type1
cl11	Lung	Blood_vessel,Lymph_vessel,Muscle_cells,Fibroblast,Alveolar_Type2
cl11	Spleen	B_mantle,T_cell_dividing,NK_dividing
cl11	Oesophagus	Blood_vessel,Lymph_vessel,Stroma,Epi_basal
cl424	Lung	Lymph_vessel,Blood_vessel,Fibroblast
cl424	Oesophagus	Lymph_vessel,Blood_vessel
cl329	Lung	Ciliated,Mast_cells,T_CD4,Alveolar_Type1,T_CD8_CytT
cl329	Spleen	T_CD8_gd,DC_activated
cl329	Oesophagus	T_CD4,T_CD8,Glands_duct,NK_T_CD8_Cytotoxic,B_CD27pos
cl128	Lung	Alveolar_Type1,Alveolar_Type2
cl128	Oesophagus	Glands_mucous,Epi_stratified
cl31	Lung	Lymph_vessel,Fibroblast,Alveolar_Type1
cl31	Spleen	DC_activated,T_CD4_naive
cl31	Oesophagus	Lymph_vessel,Epi_basal
cl8	Lung	T_cells_Dividing,T_CD4,T_CD8_CytT,T_regulatory,DC_plasmacytoid
cl8	Spleen	T_CD4_reg,T_CD4_conv,T_CD8_activated,T_cell_dividing,T_CD8_CTL
cl8	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic,Mast_cell,B_CD27neg

Table C.2: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters (continued 1)

Cluster	Tissue	Cell types
cl554	Lung	B_cells,DC_plasmacytoid,DC_activated,T_cells_Dividing
cl554	Spleen	B_follicular,B_mantle,B_Hypermutation
cl554	Oesophagus	B_CD27pos,B_CD27neg,T_CD4,Dendritic_Cells,NK_T_CD8_Cytotoxic
cl425	Lung	Lymph_vessel,Blood_vessel,Muscle_cells,Fibroblast
cl425	Spleen	DC_1
cl425	Oesophagus	Lymph_vessel,Blood_vessel,Stroma,Epi_basal,Glands_duct
cl210	Lung	NK_Dividing,NK,T_cells_Dividing,DC_plasmacytoid,DC_1
cl210	Spleen	NK_CD160pos,NK_FCGR3Apos,T_CD8_gd,NK_dividing,T_CD8_MAIT
cl210	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8,T_CD4,Dendritic_Cells,B_CD27pos
cl87	Lung	T_CD4
cl87	Spleen	T_CD8_MAIT,Monocyte
cl87	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic,Mast_cell
cl47	Lung	Fibroblast,Muscle_cells,NK_Dividing,Blood_vessel,Lymph_vessel
cl47	Oesophagus	Stroma,Blood_vessel,Epi_basal,Lymph_vessel
cl222	Lung	Lymph_vessel,Blood_vessel,Fibroblast
cl222	Oesophagus	Lymph_vessel,Blood_vessel,Stroma
cl88	Lung	Blood_vessel,Lymph_vessel,Alveolar_Type1,DC_activated,Muscle_cells
cl88	Spleen	T_CD8_MAIT,T_CD4_conv,T_CD4_naive,DC_activated
cl88	Oesophagus	Blood_vessel,Lymph_vessel,Epi_basal,Glands_duct,Stroma
cl73	Lung	T_cells_Dividing,T_CD4,T_regulatory,DC_activated
cl73	Spleen	T_CD4_reg,T_CD8_MAIT,ILC,T_CD4_fh,T_CD4_conv
cl73	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic,Mast_cell,B_CD27pos
cl606	Lung	Fibroblast,Muscle_cells,DC_activated,Macrophage_MARCOpos,Macrophage_MARCOneg
cl606	Spleen	Monocyte,DC_1,DC_2,Macrophage
cl606	Oesophagus	Stroma,Mast_cell,Epi_suprabasal,Mono_macro,Lymph_vessel
cl449	Spleen	T_cell_dividing,T_CD4_conv,T_CD4_fh,B_Hypermutation,CD34_progenitor
cl449	Oesophagus	Lymph_vessel,Blood_vessel,Glands_duct
cl58	Lung	T_regulatory,T_cells_Dividing,T_CD4,T_CD8_CytT,Mast_cells
cl58	Spleen	NK_CD160pos,T_CD4_reg,T_CD8_gd,T_CD8_MAIT,T_CD8_CTL
cl58	Oesophagus	T_CD8,T_CD4,NK_T_CD8_Cytotoxic,Mast_cell,Dendritic_Cells
cl74	Lung	Alveolar_Type1
cl74	Oesophagus	Epi_basal,Glands_duct
cl147	Spleen	CD34_progenitor
cl71	Lung	T_CD8_CytT
cl71	Spleen	T_CD8_MAIT,T_CD8_activated,T_CD8_CTL,T_CD8_gd
cl71	Oesophagus	T_CD4,T_CD8
cl616	Lung	T_CD8_CytT,NK_Dividing,NK,T_regulatory,T_cells_Dividing
cl616	Spleen	T_CD8_activated,T_CD8_MAIT,T_CD8_gd,NK_CD160pos,T_CD4_fh
cl616	Oesophagus	T_CD8,NK_T_CD8_Cytotoxic,T_CD4
cl179	Lung	NK_Dividing,NK,T_cells_Dividing
cl179	Spleen	NK_dividing,NK_CD160pos,T_CD8_gd,NK_FCGR3Apos,ILC
cl179	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8,Epi_dividing,T_CD4,Mast_cell
cl34	Lung	Fibroblast,Muscle_cells,Monocyte
cl34	Spleen	Monocyte,T_CD8_CTL
cl34	Oesophagus	Stroma,Lymph_vessel,Dendritic_Cells,Epi_basal,Mono_macro
cl271	Lung	Fibroblast,Muscle_cells,Lymph_vessel,Blood_vessel
cl271	Oesophagus	Stroma,Lymph_vessel,Epi_basal,Blood_vessel,Epi_suprabasal
cl172	Lung	Lymph_vessel,Blood_vessel,Fibroblast
cl172	Oesophagus	Lymph_vessel,Blood_vessel,Stroma,Epi_basal,Epi_suprabasal
cl435	Lung	Macrophage_MARCOneg,Macrophage_MARCOpos
cl435	Spleen	Macrophage,DC_2,Monocyte
cl79	Lung	NK_Dividing,NK,T_CD8_CytT,T_cells_Dividing,T_regulatory
cl79	Spleen	T_CD8_activated,NK_dividing,T_CD8_gd,T_CD8_CTL,NK_CD160pos
cl79	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8,T_CD4,Epi_dividing,Mono_macro
cl36	Lung	Blood_vessel,DC_activated,DC_Monocyte_Dividing,DC_plasmacytoid,Macrophage_MARCOpos
cl36	Spleen	DC_2,DC_activated,DC_1,B_follicular,Macrophage
cl36	Oesophagus	Blood_vessel,Dendritic_Cells,Mono_macro,B_CD27pos,B_CD27neg
cl505	Spleen	Monocyte
cl404	Lung	Muscle_cells,Fibroblast
cl404	Oesophagus	Epi_suprabasal

Table C.3: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters (continued 2)

Cluster	Tissue	Cell types
cl464	Lung	Macrophage_MARCOneg
cl464	Spleen	CD34_progenitor,DC_2
cl464	Oesophagus	Dendritic_Cells
cl596	Lung	T_CD4,T_cells_Dividing,T_regulatory,Mast_cells,B_cells
cl596	Spleen	T_CD8_MAIT,T_CD4_conv,T_CD4_fh,T_CD4_reg,T_CD4_naive
cl596	Oesophagus	T_CD8,NK_T_CD8_Cytotoxic,T_CD4,Dendritic_Cells,B_CD27pos
cl45	Lung	DC_2,Macrophage_MARCOneg,DC_1,DC_Monocyte_Dividing,DC_plasmacytoid
cl45	Spleen	DC_2,DC_1,DC_activated,DC_plasmacytoid,Monocyte
cl45	Oesophagus	Dendritic_Cells,Mono_macro,B_CD27pos,B_CD27neg,Blood_vessel
cl51	Lung	DC_2,Macrophage_MARCOneg,DC_1,DC_activated,DC_Monocyte_Dividing
cl51	Spleen	DC_2,DC_1,DC_activated,Monocyte,Macrophage
cl51	Oesophagus	Dendritic_Cells,Mono_macro,T_CD4,B_CD27pos,B_CD27neg
cl441	Spleen	T_CD4_naive
cl432	Lung	Alveolar_Type1,Ciliated,Alveolar_Type2
cl432	Spleen	B_mantle
cl432	Oesophagus	Glands_duct,Glands_mucous,Epi_basal
cl617	Lung	DC_2,Mast_cells,Muscle_cells
cl617	Spleen	Monocyte,T_CD8_MAIT
cl617	Oesophagus	Mast_cell,Dendritic_Cells,Epi_basal,B_CD27pos,Mono_macro
cl260	Lung	DC_plasmacytoid,Alveolar_Type1,Ciliated,Monocyte,DC_activated
cl260	Spleen	CD34_progenitor
cl260	Oesophagus	Blood_vessel,Epi_suprabasal
cl452	Lung	Monocyte
cl452	Spleen	Monocyte
cl27	Lung	Blood_vessel,Lymph_vessel,Muscle_cells
cl27	Oesophagus	Blood_vessel,Lymph_vessel,Stroma
cl64	Lung	Blood_vessel,Lymph_vessel,Fibroblast,Muscle_cells
cl64	Spleen	T_CD4_conv
cl64	Oesophagus	Blood_vessel,Lymph_vessel,Stroma,Epi_basal
cl205	Lung	T_CD8_CytT
cl205	Spleen	T_CD8_gd,T_CD8_CTL
cl252	Lung	Macrophage_MARCOpos,Macrophage_Dividing,DC_activated,DC_Monocyte_Dividing,DC_2
cl252	Spleen	DC_2,DC_1,NK_dividing,DC_activated,CD34_progenitor
cl252	Oesophagus	NK_T_CD8_Cytotoxic,T_CD4,Mast_cell,Mono_macro,T_CD8
cl76	Spleen	Monocyte
cl508	Lung	T_CD8_CytT,T_CD4,T_regulatory,NK,NK_Dividing
cl508	Spleen	T_CD8_CTL,T_CD8_MAIT,T_CD8_activated,T_CD8_gd,T_CD4_fh
cl508	Oesophagus	T_CD8,T_CD4,NK_T_CD8_Cytotoxic
cl621	Lung	T_cells_Dividing,T_CD4,T_regulatory,DC_activated
cl621	Spleen	T_CD8_MAIT,T_CD4_reg,T_cell_dividing,Monocyte,T_CD4_conv
cl621	Oesophagus	T_CD4,T_CD8,Dendritic_Cells,NK_T_CD8_Cytotoxic,Mast_cell
cl512	Lung	Monocyte,Macrophage_MARCOneg,Macrophage_MARCOpos,DC_1,DC_2
cl512	Spleen	Monocyte,DC_2,Macrophage,DC_activated,DC_1
cl512	Oesophagus	Mono_macro,Dendritic_Cells,B_CD27pos,B_CD27neg,T_CD4
cl70	Lung	DC_Monocyte_Dividing,DC_1,Macrophage_Dividing,DC_activated,Macrophage_MARCOpos
cl70	Spleen	DC_1,DC_2,DC_activated,B_follicular,B_mantle
cl70	Oesophagus	Dendritic_Cells,Blood_vessel,Mono_macro,B_CD27pos,B_CD27neg
cl568	Lung	Ciliated
cl340	Lung	Fibroblast,Lymph_vessel
cl340	Spleen	DC_plasmacytoid
cl340	Oesophagus	Glands_mucous,Stroma,Epi_basal,Lymph_vessel
cl57	Lung	Macrophage_MARCOneg,DC_2,DC_Monocyte_Dividing,DC_activated,DC_1
cl57	Spleen	DC_2,Monocyte,DC_1,DC_activated,DC_plasmacytoid
cl57	Oesophagus	Dendritic_Cells,Mono_macro,Lymph_vessel,Mast_cell,Blood_vessel
cl491	Lung	Macrophage_Dividing,DC_Monocyte_Dividing,Macrophage_MARCOpos,DC_activated,DC_2
cl491	Spleen	DC_2,DC_1,Monocyte,Macrophage,DC_activated
cl491	Oesophagus	Dendritic_Cells,Mono_macro,B_CD27neg,T_CD4,B_CD27pos

Table C.4: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters (continued 3)

Cluster	Tissue	Cell types
cl100	Lung	Blood_vessel,DC_plasmacytoid,Lymph_vessel,Muscle_cells,DC_2
cl100	Spleen	DC_plasmacytoid,DC_2,B_follicular,B_mantle,DC_1
cl100	Oesophagus	Blood_vessel,Dendritic_Cells,Lymph_vessel,Stroma,Mono_macro
cl46	Lung	DC_activated,DC_1,DC_Monocyte_Dividing,Macrophage_MARCOneg,Macrophage_MARCOpos
cl46	Spleen	DC_activated,DC_2,DC_1,B_follicular,B_Hypermutation
cl46	Oesophagus	Dendritic_Cells,Blood_vessel,Mono_macro,B_CD27neg,B_CD27pos
cl44	Lung	DC_2,Macrophage_MARCOneg,DC_Monocyte_Dividing,Macrophage_MARCOpos,Monocyte
cl44	Spleen	Monocyte,DC_2,DC_activated,Macrophage,T_CD4_conv
cl44	Oesophagus	Mono_macro,Dendritic_Cells,T_CD8,B_CD27neg,NK_T_CD8_Cytotoxic
cl503	Lung	DC_2,Macrophage_MARCOneg,Monocyte,DC_activated
cl503	Spleen	Monocyte,Macrophage
cl503	Oesophagus	Epi_basal,Blood_vessel,Mono_macro,Glands_duct
cl25	Lung	Macrophage_MARCOneg,Macrophage_MARCOpos, DC_2,Macrophage_Dividing,DC_Monocyte_Dividing
cl25	Spleen	DC_2,Monocyte,Macrophage,DC_1,DC_plasmacytoid
cl25	Oesophagus	Mono_macro,Dendritic_Cells,Mast_cell,Glands_duct,Lymph_vessel
cl93	Lung	Monocyte,Macrophage_MARCOpos
cl93	Spleen	Monocyte
cl485	Lung	Plasma_cells,DC_1
cl485	Spleen	Plasma_IgG,Plasma_IgM,Monocyte
cl577	Lung	Alveolar_Type2,Alveolar_Type1
cl577	Spleen	B_follicular
cl577	Oesophagus	Epi_upper
cl316	Lung	Alveolar_Type1
cl316	Oesophagus	Glands_mucous,Glands_duct,Epi_upper,Epi_basal
cl220	Lung	Ciliated
cl611	Lung	T_CD4,T_regulatory,DC_activated,T_cells_Dividing,DC_2
cl611	Spleen	T_CD8_MAIT,T_CD4_conv,ILC,T_CD4_fh,T_cell_dividing
cl611	Oesophagus	NK_T_CD8_Cytotoxic,T_CD4,T_CD8
cl242	Lung	Fibroblast
cl242	Oesophagus	Stroma,Epi_basal
cl458	Lung	NK_Dividing,NK,T_CD8_CytT
cl458	Spleen	T_CD8_CTL,NK_FCGR3Apos,NK_CD160pos,NK_dividing,T_CD8_MAIT
cl458	Oesophagus	T_CD8,NK_T_CD8_Cytotoxic,T_CD4
cl417	Lung	Lymph_vessel,Alveolar_Type1
cl417	Spleen	DC_1
cl417	Oesophagus	Glands_duct,Glands_mucous
cl401	Lung	DC_2,DC_activated
cl401	Oesophagus	Glands_mucous,Glands_duct
cl581	Lung	Alveolar_Type2,Alveolar_Type1,Ciliated
cl581	Oesophagus	Epi_basal
cl592	Lung	DC_plasmacytoid
cl592	Spleen	B_follicular,B_mantle
cl592	Oesophagus	B_CD27pos,B_CD27neg
cl376	Lung	Muscle_cells,Fibroblast,Ciliated
cl376	Oesophagus	Stroma,Lymph_vessel,Mast_cell
cl519	Lung	T_regulatory,T_CD4,T_cells_Dividing
cl519	Spleen	T_CD8_MAIT,T_CD4_fh,T_CD4_conv,T_CD4_reg
cl519	Oesophagus	T_CD4,NK_T_CD8_Cytotoxic,T_CD8,Mast_cell
cl35	Lung	Macrophage_MARCOpos,DC_Monocyte_Dividing,DC_1, Macrophage_Dividing,Macrophage_MARCOneg
cl35	Spleen	DC_1,DC_2,DC_activated,Monocyte,B_mantle
cl35	Oesophagus	Mono_macro,Dendritic_Cells,B_CD27neg,Glands_duct,B_CD27pos
cl446	Lung	T_CD4,T_CD8_CytT,T_regulatory
cl446	Spleen	T_CD4_fh,T_CD4_reg,T_CD8_MAIT,T_CD4_naive
cl446	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic
cl219	Lung	Blood_vessel,Muscle_cells,Lymph_vessel,Alveolar_Type1,Fibroblast
cl219	Oesophagus	Blood_vessel,Lymph_vessel,Stroma,Epi_basal

Table C.5: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters (continued 4)

Cluster	Tissue	Cell types
cl496	Lung	T_CD4
cl496	Spleen	T_CD4_conv,T_CD4_naive
cl69	Lung	DC_plasmacytoid,Plasma_cells,B_cells,DC_1,Macrophage_MARCOneg
cl69	Spleen	B_follicular,Plasma_IgM,DC_plasmacytoid,Plasmablast,B_mantle
cl69	Oesophagus	B_CD27neg,B_CD27pos,Blood_vessel,Glands_duct,Dendritic_Cells
cl134	Spleen	CD34_progenitor
cl28	Lung	T_cells_Dividing,NK_Dividing
cl28	Spleen	NK_dividing,T_cell_dividing
cl14	Lung	T_cells_Dividing,T_regulatory,T_CD4,T_CD8_CytT,NK_Dividing
cl14	Spleen	T_CD8_CTL,T_CD4_reg,T_CD4_fh,T_cell_dividing,T_CD4_conv
cl14	Oesophagus	T_CD8,T_CD4,NK_T_CD8_Cytotoxic,Dendritic_Cells,B_CD27pos
cl40	Lung	DC_plasmacytoid,B_cells,DC_1,DC_activated,DC_2
cl40	Spleen	B_follicular,B_mantle,DC_plasmacytoid,DC_activated
cl40	Oesophagus	B_CD27neg,B_CD27pos,Dendritic_Cells
cl509	Lung	DC_1,DC_2,Macrophage_MARCOneg
cl509	Spleen	B_follicular,B_mantle
cl509	Oesophagus	Mono_macro,B_CD27pos,Dendritic_Cells,B_CD27neg
cl113	Lung	DC_plasmacytoid,DC_activated
cl113	Spleen	B_mantle,B_follicular
cl507	Lung	T_regulatory,Ciliated
cl507	Spleen	T_CD4_naive,T_CD4_conv
cl102	Lung	Fibroblast,Muscle_cells,Lymph_vessel
cl102	Oesophagus	Stroma,Epi_basal,Blood_vessel,Lymph_vessel
cl590	Lung	T_CD4,T_cells_Dividing,T_regulatory
cl590	Spleen	T_CD4_fh,T_CD4_reg,T_CD4_naive,T_CD4_conv
cl590	Oesophagus	NK_T_CD8_Cytotoxic,T_CD4
cl422	Lung	Macrophage_MARCOneg,DC_2,DC_Monocyte_Dividing, Macrophage_MARCOpos,Macrophage_Dividing
cl422	Spleen	DC_2,CD34_progenitor,Unknown,NK_FCGR3Apos,Monocyte
cl422	Oesophagus	Dendritic_Cells,B_CD27neg,NK_T_CD8_Cytotoxic,Mono_macro,T_CD8
cl106	Lung	DC_1,B_cells,DC_activated,DC_2
cl106	Spleen	B_follicular,B_mantle,B_Hypermutation
cl106	Oesophagus	B_CD27pos,B_CD27neg,Dendritic_Cells,Blood_vessel,T_CD4
cl183	Lung	Mast_cells,T_CD4,DC_Monocyte_Dividing,T_cells_Dividing,DC_plasmacytoid
cl183	Spleen	T_CD8_MAIT,Monocyte,T_CD4_reg,CD34_progenitor,B_follicular
cl183	Oesophagus	Mast_cell,Dendritic_Cells,NK_T_CD8_Cytotoxic,T_CD8,T_CD4
cl68	Lung	T_CD4,T_cells_Dividing,T_regulatory,DC_1,Mast_cells
cl68	Spleen	T_CD4_naive,T_CD4_fh,T_CD4_conv,ILC,T_CD4_reg
cl68	Oesophagus	T_CD4,NK_T_CD8_Cytotoxic,T_CD8,Dendritic_Cells,B_CD27pos
cl192	Lung	Monocyte,Macrophage_MARCOpos,DC_2,DC_Monocyte_Dividing,Macrophage_MARCOneg
cl192	Spleen	Monocyte,DC_2,Macrophage
cl192	Oesophagus	Mono_macro,Dendritic_Cells,Mast_cell,B_CD27pos,T_CD4
cl16	Lung	DC_activated,Blood_vessel,Plasma_cells,DC_1,DC_Monocyte_Dividing
cl16	Spleen	Plasma_IgG,Plasmablast,Plasma_IgM,B_follicular,B_mantle
cl16	Oesophagus	Blood_vessel,B_CD27neg,B_CD27pos,Lymph_vessel,Dendritic_Cells
cl608	Lung	DC_Monocyte_Dividing,T_cells_Dividing,NK_Dividing,DC_activated,DC_1
cl608	Spleen	B_Hypermutation,T_cell_dividing,B_follicular,DC_2,NK_dividing
cl608	Oesophagus	B_CD27pos,B_CD27neg,T_CD4,Epi_dividing,T_CD8
cl470	Lung	NK_Dividing,NK,T_CD8_CytT
cl470	Spleen	NK_FCGR3Apos,T_CD8_CTL,T_CD8_MAIT,NK_dividing,NK_CD160pos
cl470	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8
cl53	Spleen	Platelet,CD34_progenitor
cl479	Oesophagus	B_CD27neg
cl124	Lung	Plasma_cells
cl124	Spleen	Plasma_IgM,Plasma_IgG,Plasmablast
cl124	Oesophagus	B_CD27pos
cl131	Lung	Macrophage_MARCOpos,Macrophage_Dividing,Macrophage_MARCOneg, DC_2,DC_Monocyte_Dividing
cl131	Spleen	Monocyte,DC_2,DC_1,Macrophage,B_follicular
cl131	Oesophagus	Mono_macro,Glands_duct,Dendritic_Cells,B_CD27pos,Blood_vessel

Table C.6: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters (continued 5)

Cluster	Tissue	Cell types
cl152	Lung	Ciliated,Alveolar_Type1,Alveolar_Type2
cl152	Oesophagus	Glands_mucous,Glands_duct
cl306	Lung	Muscle_cells,Blood_vessel,Fibroblast,Lymph_vessel,Alveolar_Type1
cl306	Oesophagus	Stroma,Blood_vessel,Epi_basal,Lymph_vessel,Epi_suprabasal
cl115	Lung	Monocyte,Macrophage_Dividing,Macrophage_MARCOpos,DC_Monocyte_Dividing
cl115	Spleen	Monocyte,NK_CD160pos,T_CD8_CTL
cl115	Oesophagus	Mono_macro,Dendritic_Cells,T_CD8,NK_T_CD8_Cytotoxic,Mast_cell
cl601	Lung	B_cells,DC_1,DC_2,DC_plasmacytoid,DC_activated
cl601	Spleen	B_follicular,B_mantle,DC_activated,T_CD8_gd
cl601	Oesophagus	B_CD27pos,B_CD27neg,Dendritic_Cells,Mono_macro,T_CD4
cl307	Lung	Alveolar_Type2,Alveolar_Type1,DC_activated,Ciliated,Monocyte
cl307	Spleen	Plasmablast,DC_activated,DC_1,Plasma_IgG,Plasma_IgM
cl307	Oesophagus	Mast_cell
cl426	Lung	Lymph_vessel,Blood_vessel,DC_activated,Fibroblast
cl426	Spleen	T_cell_dividing,DC_1,DC_activated
cl426	Oesophagus	Lymph_vessel,Blood_vessel,Stroma
cl428	Lung	Fibroblast,Muscle_cells,Lymph_vessel,Blood_vessel,Alveolar_Type2
cl428	Spleen	T_CD4_fh
cl428	Oesophagus	Stroma,Lymph_vessel,Epi_suprabasal,Blood_vessel,Epi_basal
cl212	Lung	DC_1,DC_Monocyte_Dividing,DC_2,DC_activated,T_cells_Dividing
cl212	Spleen	DC_1,DC_2,DC_activated,B_follicular,NK_dividing
cl212	Oesophagus	Dendritic_Cells,B_CD27pos,B_CD27neg,Mono_macro,NK_T_CD8_Cytotoxic
cl85	Lung	Muscle_cells,Blood_vessel,Fibroblast,Lymph_vessel,Alveolar_Type1
cl85	Spleen	B_follicular,T_CD4_conv,B_mantle
cl85	Oesophagus	Blood_vessel,Stroma,Lymph_vessel,Epi_basal,Epi_suprabasal
cl133	Lung	Macrophage_MARCOpos,Macrophage_Dividing,Mast_cells,Alveolar_Type1,Alveolar_Type2
cl133	Oesophagus	Dendritic_Cells,Glands_duct,Mono_macro,Epi_basal
cl414	Lung	DC_Monocyte_Dividing,T_cells_Dividing,NK_Dividing,DC_1,DC_2
cl414	Spleen	NK_dividing,T_cell_dividing,B_Hypermutation,DC_1,DC_2
cl414	Oesophagus	Epi_dividing,Dendritic_Cells,Mono_macro,B_CD27pos,NK_T_CD8_Cytotoxic
cl516	Lung	T_regulatory,T_CD4
cl516	Spleen	T_CD4_reg,T_CD4_naive,T_CD4_fh,T_CD4_conv
cl516	Oesophagus	T_CD4,NK_T_CD8_Cytotoxic,T_CD8
cl379	Lung	Alveolar_Type2,Alveolar_Type1,Macrophage_MARCOneg, Macrophage_Dividing,Macrophage_MARCOpos
cl379	Spleen	Macrophage
cl379	Oesophagus	Dendritic_Cells,Mono_macro,Glands_duct
cl83	Lung	Fibroblast,Lymph_vessel,Blood_vessel
cl83	Oesophagus	Stroma,Lymph_vessel,Blood_vessel
cl412	Lung	Macrophage_MARCOneg,DC_2,DC_1,DC_Monocyte_Dividing,DC_plasmacytoid
cl412	Spleen	DC_2,DC_1,Monocyte,DC_plasmacytoid,B_follicular
cl412	Oesophagus	Dendritic_Cells,Mono_macro,NK_T_CD8_Cytotoxic,B_CD27neg,T_CD4
cl285	Lung	T_cells_Dividing,DC_plasmacytoid,Plasma_cells
cl285	Spleen	DC_activated,Plasmablast,DC_plasmacytoid,DC_1
cl285	Oesophagus	T_CD8,T_CD4,NK_T_CD8_Cytotoxic,B_CD27pos
cl281	Lung	Fibroblast,Lymph_vessel,Muscle_cells,Macrophage_MARCOpos,Blood_vessel
cl281	Spleen	DC_1
cl281	Oesophagus	Epi_basal,Stroma,Glands_duct,Blood_vessel,Lymph_vessel
cl382	Lung	Muscle_cells,Fibroblast,Alveolar_Type1,Lymph_vessel,Blood_vessel
cl382	Spleen	T_CD8_CTL,T_cell_dividing,NK_dividing,T_CD8_activated,T_CD4_reg
cl382	Oesophagus	Stroma,Glands_duct,Epi_basal,Epi_suprabasal,Blood_vessel
cl97	Lung	Muscle_cells,Fibroblast,DC_Monocyte_Dividing,T_cells_Dividing,Blood_vessel
cl97	Spleen	T_cell_dividing,NK_dividing,B_Hypermutation,Plasmablast
cl97	Oesophagus	Stroma,Epi_dividing,Epi_upper,Epi_suprabasal
cl595	Lung	DC_plasmacytoid,Mast_cells,DC_2,DC_activated,T_cells_Dividing
cl595	Spleen	T_CD8_gd,Plasma_IgG,Plasma_IgM
cl595	Oesophagus	Mast_cell,T_CD4,NK_T_CD8_Cytotoxic,Glands_mucous,B_CD27pos

Table C.7: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters (continued 6)

Cluster	Tissue	Cell types
cl75	Lung	Plasma_cells,B_cells
cl75	Spleen	Plasma_IgG,Plasma_IgM,Plasmablast
cl75	Oesophagus	B_CD27pos
cl268	Lung	Fibroblast,Muscle_cells
cl268	Oesophagus	Stroma
cl620	Spleen	Plasmablast
cl273	Lung	NK_Dividing,T_CD8_CytT,T_cells_Dividing,NK,T_CD4
cl273	Spleen	T_CD8_gd,NK_CD160pos,T_CD8_MAIT,T_CD8_activated,T_CD8_CTL
cl273	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8,T_CD4,Dendritic_Cells,B_CD27neg
cl543	Lung	Plasma_cells
cl543	Spleen	Plasma_IgM,Plasma_IgG
cl13	Lung	Muscle_cells,Fibroblast,DC_activated,DC_1,Blood_vessel
cl13	Spleen	T_CD8_gd
cl13	Oesophagus	Stroma,Glands_duct,Blood_vessel,Lymph_vessel,Glands_mucous
cl546	Lung	B_cells,DC_plasmacytoid,DC_1,Macrophage_MARCOneg,Plasma_cells
cl546	Spleen	B_mantle,B_follicular,DC_activated
cl546	Oesophagus	B_CD27neg,B_CD27pos,Dendritic_Cells,Mono_macro,Blood_vessel
cl23	Lung	DC_2,Monocyte,DC_1,DC_activated,DC_Monocyte_Dividing
cl23	Spleen	DC_2,Monocyte,DC_1,DC_activated,B_mantle
cl23	Oesophagus	Dendritic_Cells,Mono_macro,B_CD27pos,B_CD27neg,T_CD4
cl490	Lung	NK,NK_Dividing,Ciliated
cl490	Spleen	NK_FCGR3Apos,NK_CD160pos,NK_dividing
cl490	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8
cl497	Lung	DC_2,DC_activated,Macrophage_MARCOneg,DC_1,DC_plasmacytoid
cl497	Spleen	B_follicular,DC_activated
cl497	Oesophagus	Dendritic_Cells,Mono_macro,Blood_vessel,Glands_duct
cl438	Lung	DC_1,DC_Monocyte_Dividing,T_CD4,DC_2,Macrophage_MARCOneg
cl438	Spleen	CD34_progenitor,DC_1,DC_plasmacytoid,ILC,B_Hypermutation
cl438	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic,B_CD27neg,B_CD27pos
cl495	Lung	NK_Dividing,T_cells_Dividing
cl495	Oesophagus	Epi_dividing
cl110	Lung	Macrophage_MARCOpos,Macrophage_MARCOneg,Lymph_vessel, Mast_cells,Macrophage_Dividing
cl110	Spleen	Macrophage,Monocyte,DC_plasmacytoid,DC_2,T_CD4_conv
cl110	Oesophagus	Mono_macro,Stroma,Glands_duct,Lymph_vessel,Mast_cell
cl269	Lung	T_cells_Dividing,T_CD4,T_CD8_CytT,NK_Dividing,NK
cl269	Spleen	T_CD8_CTL,T_CD8_MAIT,T_CD8_activated,T_CD8_gd,NK_CD160pos
cl269	Oesophagus	T_CD8,NK_T_CD8_Cytotoxic,T_CD4,Dendritic_Cells,Mast_cell
cl127	Lung	Macrophage_MARCOpos,Macrophage_Dividing,DC_2,Monocyte,Macrophage_MARCOneg
cl127	Spleen	Macrophage,DC_2,Monocyte,B_follicular,DC_1
cl127	Oesophagus	Mono_macro,Dendritic_Cells,B_CD27neg,B_CD27pos,Blood_vessel
cl311	Lung	Muscle_cells,Fibroblast,Blood_vessel,Lymph_vessel,Alveolar_Type1
cl311	Spleen	T_CD8_CTL,T_CD4_fh,T_CD8_activated,T_CD4_reg,T_CD4_conv
cl311	Oesophagus	Stroma,Lymph_vessel,Blood_vessel,Epi_suprabasal,Epi_basal
cl118	Lung	Macrophage_MARCOpos,Macrophage_Dividing,Macrophage_MARCOneg,DC_2,Monocyte
cl118	Spleen	Monocyte,DC_2,Macrophage
cl118	Oesophagus	Mono_macro,Dendritic_Cells
cl77	Lung	DC_plasmacytoid,DC_activated
cl77	Spleen	DC_plasmacytoid,B_follicular,DC_activated
cl77	Oesophagus	B_CD27pos,Dendritic_Cells,Blood_vessel
cl472	Lung	NK_Dividing,NK,T_CD8_CytT
cl472	Spleen	T_CD8_CTL,NK_FCGR3Apos,T_CD8_MAIT,NK_CD160pos,T_CD8_activated
cl610	Lung	T_CD4,T_cells_Dividing,T_regulatory
cl610	Spleen	T_CD4_fh,T_CD4_reg,T_CD4_conv,T_CD4_naive,T_CD8_activated
cl610	Oesophagus	T_CD8,T_CD4,NK_T_CD8_Cytotoxic
cl180	Lung	Lymph_vessel,Blood_vessel,Fibroblast
cl180	Oesophagus	Lymph_vessel,Blood_vessel,Stroma,Glands_duct,Epi_basal
cl251	Lung	DC_1,Plasma_cells,Monocyte,DC_activated
cl251	Spleen	Plasma_IgM,B_follicular,Plasma_IgG
cl251	Oesophagus	T_CD8,T_CD4,B_CD27neg

Table C.8: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters (continued 7)

Cluster	Tissue	Cell types
cl380	Lung	Fibroblast,DC_2,DC_1
cl380	Spleen	Monocyte
cl380	Oesophagus	Stroma,NK_T_CD8_Cytotoxic,Epi_basal,Lymph_vessel,Glands_duct
cl591	Lung	Plasma_cells
cl591	Spleen	Plasma_IgM,Plasma_IgG,Plasmablast
cl591	Oesophagus	B_CD27pos
cl266	Lung	Monocyte,DC_2,Macrophage_MARCOneg,DC_1,DC_activated
cl266	Spleen	Monocyte,DC_2,DC_plasmacytoid,DC_1,Macrophage
cl266	Oesophagus	Dendritic_Cells,Mono_macro,B_CD27pos,NK_T_CD8_Cytotoxic,T_CD4
cl510	Lung	Macrophage_Dividing,DC_Monocyte_Dividing
cl510	Spleen	Monocyte
cl510	Oesophagus	Mono_macro
cl101	Lung	DC_plasmacytoid,DC_activated,Mast_cells,T_cells_Dividing,Plasma_cells
cl101	Spleen	B_follicular,DC_activated,Plasmablast,T_CD8_activated,T_CD8_gd
cl101	Oesophagus	Dendritic_Cells,B_CD27pos,Blood_vessel,Glands_duct,B_CD27neg
cl24	Lung	Macrophage_MARCOneg,DC_Monocyte_Dividing,DC_2,Monocyte,Macrophage_MARCOpos
cl24	Spleen	Monocyte,DC_2,DC_1,Macrophage,B_mantle
cl24	Oesophagus	Dendritic_Cells,Mono_macro,T_CD4,B_CD27pos,B_CD27neg
cl473	Lung	T_CD4,T_regulatory,T_cells_Dividing,T_CD8_CytT,DC_activated
cl473	Spleen	T_CD8_MAIT,T_CD8_CTL,T_CD4_conv,T_CD4_naive,T_CD4_fh
cl473	Oesophagus	T_CD8,T_CD4,NK_T_CD8_Cytotoxic
cl7	Lung	Muscle_cells,Fibroblast,Lymph_vessel,Blood_vessel,Alveolar_Type1
cl7	Spleen	T_CD8_CTL
cl7	Oesophagus	Stroma,Blood_vessel,Lymph_vessel,Glands_duct,Epi_basal
cl618	Lung	T_cells_Dividing,T_regulatory,T_CD4,B_cells,DC_Monocyte_Dividing
cl618	Spleen	Plasmablast
cl618	Oesophagus	B_CD27pos,T_CD8,NK_T_CD8_Cytotoxic,B_CD27neg
cl155	Lung	T_cells_Dividing,DC_Monocyte_Dividing,NK_Dividing,Macrophage_Dividing,B_cells
cl155	Spleen	T_cell_dividing,B_Hypermutation,Plasmablast,NK_dividing,B_follicular
cl155	Oesophagus	Epi_dividing,B_CD27pos,B_CD27neg
cl67	Lung	Lymph_vessel,Blood_vessel,Fibroblast,Muscle_cells,DC_Monocyte_Dividing
cl67	Spleen	DC_1,NK_dividing,T_cell_dividing,B_Hypermutation,DC_2
cl67	Oesophagus	Lymph_vessel,Blood_vessel,Stroma,Glands_duct,Epi_basal
cl0	Lung	Blood_vessel,DC_Monocyte_Dividing,Lymph_vessel,DC_1,DC_plasmacytoid
cl0	Spleen	NK_dividing,B_Hypermutation,T_cell_dividing,CD34_progenitor,DC_1
cl0	Oesophagus	Blood_vessel,Lymph_vessel,Dendritic_Cells,Mono_macro,B_CD27pos
cl524	Lung	NK_Dividing,T_cells_Dividing,DC_Monocyte_Dividing,NK,Macrophage_Dividing
cl524	Spleen	NK_dividing,T_cell_dividing,NK_CD160pos,B_Hypermutation,NK_FCGR3Apos
cl524	Oesophagus	Epi_dividing,NK_T_CD8_Cytotoxic,T_CD8,T_CD4
cl514	Lung	T_CD8_CytT,T_cells_Dividing,T_CD4,T_regulatory
cl514	Spleen	T_CD8_activated,T_CD8_gd,T_CD8_MAIT,T_CD8_CTL,T_CD4_reg
cl514	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic
cl474	Lung	Mast_cells,DC_1
cl474	Spleen	CD34_progenitor
cl474	Oesophagus	Mast_cell,Mono_macro,Dendritic_Cells,NK_T_CD8_Cytotoxic,T_CD8
cl619	Lung	T_cells_Dividing,T_CD4,Monocyte,T_regulatory,Mast_cells
cl619	Spleen	T_CD8_MAIT,T_CD4_reg,T_CD4_fh,T_CD4_conv,ILC
cl619	Oesophagus	NK_T_CD8_Cytotoxic,T_CD4,T_CD8,Dendritic_Cells,Mast_cell
cl486	Lung	Monocyte,Macrophage_MARCOneg,DC_2,Macrophage_MARCOpos,T_CD4
cl486	Spleen	Monocyte,DC_2
cl486	Oesophagus	Mono_macro,Dendritic_Cells,B_CD27neg,NK_T_CD8_Cytotoxic,B_CD27pos
cl502	Lung	T_CD8_CytT,T_cells_Dividing,NK_Dividing,NK
cl502	Spleen	T_CD8_CTL,T_CD8_activated,T_CD8_MAIT,T_CD8_gd,T_CD4_reg
cl502	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic
cl501	Lung	NK,NK_Dividing
cl501	Spleen	NK_CD160pos,NK_FCGR3Apos,NK_dividing,T_CD8_gd
cl501	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8
cl55	Lung	Alveolar_Type1,Lymph_vessel,Blood_vessel
cl55	Oesophagus	Epi_basal,Glands_duct,Epi_suprabasal,Glands_mucous,Blood_vessel

Table C.9: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters (continued 8)

Cluster	Tissue	Cell types
cl529	Lung	T_CD8_CytT,NK,NK_Dividing,Monocyte,T_CD4
cl529	Spleen	T_CD8_CTL,T_CD8_MAIT,T_CD8_activated,NK_FCGR3Apos,NK_CD160pos
cl529	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8,T_CD4,Dendritic_Cells,Mono_macro
cl26	Lung	Fibroblast,Muscle_cells
cl26	Oesophagus	Stroma,Epi_suprabasal
cl60	Lung	NK_Dividing,NK,T_CD8_CytT,Macrophage_MARCOpos,DC_Monocyte_Dividing
cl60	Spleen	T_CD8_CTL,NK_FCGR3Apos,NK_CD160pos,T_CD8_gd,T_CD8_MAIT
cl60	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8,T_CD4,Dendritic_Cells,Mono_macro
cl236	Lung	Alveolar_Type1,Alveolar_Type2,Ciliated,Muscle_cells,Lymph_vessel
cl236	Spleen	T_CD4_conv,T_CD8_MAIT,T_cell_dividing,NK_dividing,T_CD8_CTL
cl236	Oesophagus	Epi_upper,Glands_duct,Epi_basal,Glands_mucous,Epi_stratified
cl538	Lung	NK_Dividing,T_cells_Dividing
cl538	Spleen	Unknown,NK_dividing
cl538	Oesophagus	Epi_dividing
cl238	Lung	Mast_cells,T_cells_Dividing,DC_plasmacytoid,T_CD4,T_regulatory
cl238	Spleen	Plasma_IgM,T_CD4_reg,T_CD8_MAIT
cl238	Oesophagus	T_CD4,T_CD8,Dendritic_Cells,NK_T_CD8_Cytotoxic,Glands_mucous
cl439	Spleen	B_mantle
cl465	Lung	Alveolar_Type2
cl465	Spleen	T_cell_dividing,Unknown,B_Hypermutation
cl447	Lung	T_CD4
cl447	Spleen	Unknown
cl283	Lung	T_regulatory,T_CD4,T_cells_Dividing,T_CD8_CytT,DC_plasmacytoid
cl283	Spleen	T_CD4_reg,T_CD4_conv,T_CD8_MAIT,T_CD8_activated,T_CD4_fh
cl283	Oesophagus	T_CD4,T_CD8
cl52	Lung	Fibroblast,Muscle_cells,Lymph_vessel,Blood_vessel,Alveolar_Type1
cl52	Spleen	T_CD8_gd,T_CD8_CTL,T_CD4_conv,T_CD8_MAIT
cl52	Oesophagus	Stroma,Lymph_vessel,Epi_basal,Blood_vessel,Epi_suprabasal
cl615	Lung	T_regulatory,T_CD4,T_cells_Dividing,T_CD8_CytT
cl615	Spleen	T_CD4_fh,T_CD4_reg,T_CD8_activated,T_CD8_gd
cl615	Oesophagus	T_CD4,T_CD8,NK_T_CD8_Cytotoxic,B_CD27neg
cl237	Lung	Plasma_cells,DC_activated,T_regulatory,Mast_cells
cl237	Spleen	Plasma_IgM,Plasma_IgG,B_follicular
cl237	Oesophagus	T_CD8
cl443	Lung	Monocyte,DC_2,DC_Monocyte_Dividing,DC_1,DC_activated
cl443	Spleen	Macrophage,T_CD8_gd,Monocyte,DC_activated,NK_CD160pos
cl443	Oesophagus	Mono_macro,Dendritic_Cells,NK_T_CD8_Cytotoxic,B_CD27neg,T_CD4
cl626	Lung	Plasma_cells,DC_2
cl626	Spleen	Plasma_IgM,T_CD8_gd,Plasmablast
cl547	Lung	Plasma_cells,Alveolar_Type2,Fibroblast,DC_Monocyte_Dividing,DC_plasmacytoid
cl547	Spleen	Plasma_IgM,Plasma_IgG,Plasmablast,DC_plasmacytoid,B_follicular
cl547	Oesophagus	Glands_mucous,B_CD27pos,Dendritic_Cells,T_CD8
cl275	Lung	Blood_vessel,Muscle_cells,Lymph_vessel,Fibroblast
cl275	Spleen	B_follicular,CD34_progenitor
cl275	Oesophagus	Blood_vessel,Lymph_vessel,Stroma,Epi_basal,Epi_suprabasal
cl185	Lung	T_cells_Dividing,T_CD4,T_regulatory
cl185	Spleen	T_CD4_reg,T_CD4_conv,T_CD8_activated,T_CD8_MAIT,T_CD4_fh
cl185	Oesophagus	T_CD8,T_CD4,NK_T_CD8_Cytotoxic
cl515	Lung	Monocyte,Macrophage_MARCOpos
cl515	Spleen	Monocyte,Macrophage,DC_2,T_CD8_MAIT,NK_CD160pos
cl515	Oesophagus	Glands_duct,Mono_macro
cl148	Lung	Lymph_vessel,Blood_vessel,Fibroblast,Muscle_cells,Alveolar_Type1
cl148	Spleen	T_CD8_gd,DC_1
cl148	Oesophagus	Lymph_vessel,Stroma,Blood_vessel,Epi_basal,Glands_mucous
cl569	Lung	Plasma_cells,B_cells,Alveolar_Type2
cl569	Spleen	Plasma_IgM,Plasma_IgG,Plasmablast,B_follicular
cl569	Oesophagus	B_CD27pos,Glands_mucous,B_CD27neg
cl282	Lung	Muscle_cells,Fibroblast,Blood_vessel,Lymph_vessel,Alveolar_Type1
cl282	Oesophagus	Stroma,Blood_vessel,Lymph_vessel,Epi_basal,Epi_suprabasal

Table C.10: Cell types from (Madissoon et al., 2019) with expression programmes enriched in *CellTypist* clusters (continued 9)

Cluster	Tissue	Cell types
cl314	Lung	Fibroblast,Muscle_cells,Lymph_vessel
cl314	Spleen	DC_1,DC_2,DC_activated
cl314	Oesophagus	Epi_suprabasal,Stroma
cl378	Lung	Fibroblast,Alveolar_Type1
cl378	Oesophagus	Stroma,Epi_basal,Epi_suprabasal
cl20	Lung	DC_2,DC_1,Monocyte,DC_Monocyte_Dividing,DC_activated
cl20	Spleen	DC_2,DC_1,Monocyte,DC_activated,CD34_progenitor
cl20	Oesophagus	Dendritic_Cells,Mono_macro,T_CD4,T_CD8,Mast_cell
cl168	Lung	T_cells_Dividing,T_CD4,T_regulatory,T_CD8_CytT,DC_Monocyte_Dividing
cl168	Spleen	T_cell_dividing,T_CD8_MAIT,T_CD4_fh,T_CD4_conv,NK_dividing
cl168	Oesophagus	NK_T_CD8_Cytotoxic,T_CD4,T_CD8,Dendritic_Cells
cl308	Lung	Alveolar_Type2,Alveolar_Type1
cl308	Oesophagus	Glands_mucous
cl304	Lung	Blood_vessel
cl304	Spleen	DC_1,NK_dividing
cl537	Lung	T_cells_Dividing,NK_Dividing,DC_Monocyte_Dividing,T_CD4,T_regulatory
cl537	Spleen	T_cell_dividing,NK_dividing,T_CD4_reg,B_Hypermutation,T_CD8_MAIT
cl537	Oesophagus	Epi_dividing,T_CD4,T_CD8,NK_T_CD8_Cytotoxic,B_CD27pos
cl613	Lung	B_cells,DC_plasmacytoid,DC_1,DC_Monocyte_Dividing,DC_activated
cl613	Spleen	B_follicular,B_mantle,B_Hypermutation
cl613	Oesophagus	B_CD27pos,B_CD27neg,Mono_macro,Dendritic_Cells
cl574	Lung	Plasma_cells
cl574	Spleen	Plasmablast,Plasma_IgM,Plasma_IgG
cl493	Spleen	Platelet
cl550	Lung	B_cells,T_CD4,T_regulatory,T_CD8_CytT,DC_activated
cl550	Spleen	T_CD8_CTL,B_follicular,B_mantle,T_CD8_MAIT,T_CD8_activated
cl550	Oesophagus	T_CD4,B_CD27neg,B_CD27pos,NK_T_CD8_Cytotoxic,T_CD8
cl492	Lung	T_cells_Dividing,NK_Dividing,DC_Monocyte_Dividing,Macrophage_Dividing,T_CD4
cl492	Spleen	NK_dividing,B_Hypermutation,CD34_progenitor,T_cell_dividing,Plasmablast
cl492	Oesophagus	Epi_dividing,B_CD27pos,B_CD27neg,NK_T_CD8_Cytotoxic
cl265	Lung	Plasma_cells
cl265	Spleen	Plasma_IgM,Plasma_IgG,Plasmablast,B_follicular,NK_dividing
cl265	Oesophagus	Glands_mucous
cl310	Lung	Plasma_cells,DC_activated,DC_1,T_regulatory,Mast_cells
cl310	Spleen	Plasma_IgG,Plasma_IgM,Plasmablast,DC_plasmacytoid,T_cell_dividing
cl310	Oesophagus	B_CD27pos,Glands_mucous,B_CD27neg,NK_T_CD8_Cytotoxic,T_CD8
cl500	Spleen	NK_FCGR3Apos,T_CD8_CTL
cl576	Lung	Alveolar_Type1,Alveolar_Type2,Macrophage_MARCOneg
cl576	Spleen	T_CD8_activated,T_CD8_CTL
cl403	Lung	DC_Monocyte_Dividing,T_cells_Dividing,NK_Dividing,Mast_cells,Monocyte
cl403	Spleen	NK_dividing,T_cell_dividing,Plasmablast,Platelet,B_Hypermutation
cl403	Oesophagus	Epi_dividing,Mast_cell,Dendritic_Cells,Lymph_vessel
cl240	Lung	NK_Dividing,DC_Monocyte_Dividing,NK,Macrophage_Dividing,T_CD8_CytT
cl240	Spleen	NK_dividing,NK_CD160pos,T_CD8_gd,NK_FCGR3Apos,B_Hypermutation
cl240	Oesophagus	NK_T_CD8_Cytotoxic,T_CD8,T_CD4,Mast_cell,Mono_macro
cl230	Lung	T_regulatory
cl549	Lung	Mast_cells

Appendix D

Publications contributed to during the PhD degree

This Appendix lists the publications to which I contributed as part of work developed during my PhD research. List updated at time of submission.

Vieira Braga F, Kar G, Berg M, Carpaij O, Polanski K, Simon L, Brouwer S, **Gomes T**, Hesse L, Jiang J, Fasouli E, Efremova M, Vento-Tormo R, Talavera-López C, Jonker M, Affleck K, Palit S, Strzelecka P, Firth H, ... Teichman, SA. (2019) A cellular census of human lungs identifies novel cell states in health and in asthma. Nature Medicine 25: 1153-1163

Miragaia, R.*, **Gomes, T.***, Chomka, A., Jardine, L., Riedel, A., Hegazy, A., Whibley, N., Tucci, A., Chen, X., Lindeman, I., Emerton G, Krausgruber T, Shields J, Haniffa M, Powrie F, and Teichmann S. (2019) Single-Cell Transcriptomics of Regulatory T Cells RevealsTrajectories of Tissue Adaptation. Immunity 50, 493-504.e7.

Lun A, Riesenfeld S, Andrews T, Dao T, **Gomes T**, and Marioni J (2019) EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biology 20:

Henriksson J, Chen X, **Gomes T**, Ullah U, Meyer K, Miragaia R, Duddy G, Pramanik J, Yusa K, Lahesmaa R, and Teichmann SA. (2019) Genome-wide CRISPR Screens in T Helper Cells Reveal Pervasive Crosstalk between Activation and Differentiation. Cell 176: 882-896.e18

Hagai T, Chen X, Miragaia R, Rostom R, **Gomes T**, Kunowska N, Henriksson J, Park J, Proserpio V, Donati G, Bossini-Castillo L, Vieira Braga F, Naamati G, Fletcher J, Stephenson E, Vegh P, Trynka G, Kondova I, Dennis M, ... Teichmann, SA. (2018) Gene expression variability across cells and species shapes innate immunity. Nature 563: 197-202

Kunz, DJ; **Gomes, T**; James, KR; Immune cell dynamics unfolded by single-cell technologies, (2018), Frontiers in immunology, 9, 1435

Pramanik J, Chen X, Kar G, Henriksson J, **Gomes T**, Park J, Natarajan K, Meyer K, Miao Z, McKenzie A, Mahata B, and Teichmann S (2018) Genome-wide analyses reveal the IRE1a-XBP1 pathway promotes T helper cell differentiation by resolving secretory stress and accelerating proliferation. Genome Medicine 10:

Miragaia, RJ; Zhang, X; **Gomes, T**; Svensson, V; Ilicic, T; Henriksson, J; Kar, G; Lönnberg, T. (2018) Single-cell RNA-sequencing resolves self-antigen expression during mTEC development, Scientific Reports, 8, 1, 685, Nature Publishing Group